



## Spotlight on...completing the count

An overview of methods for ensuring quality and completeness of your census estimates



## Introductory overview

## Introductory Overview



- **Introduce the team**
- **Agenda and aims and objectives of the day**
- **Objectives of the 2011 Census**
- **Main issues with the 2001 Census**
- **Brief overview of improvements for 2011**

## AGENDA



- **10:00 Welcome and introductory overview**
- **10:15 Introduction to coverage estimation and interactive group exercise**
- **11:10 Break**
- **11:30 Coverage estimation and group exercise continued**
- **12:30 Review of exercise and question time**
- **12:45 Lunch**
- **13:30 Overview of quality assurance**
- **13:45 Interactive exercise**
- **15:00 Review of exercise and question time**
- **15:15 Break**
- **15:35 Supplementary quality assurance**
- **15:50 Process of getting to a final estimate**
- **16:10 Wrap up session and question time**
- **16.30 End of session**

## Why are we holding these events?



- Part of our ongoing engagement with users
- Responding to UK Statistics Authority recommendations

### **Builds on recent engagement:**

- Census Regional Champion events
- Independent review of coverage estimation and quality assurance
- Presentations at British Society for Population Studies (BSPS) and other events

## Aims of tutorials



- To increase knowledge and understanding of coverage estimation and quality assurance
- To highlight improvements to the methodology for producing Census estimates since 2001
- To give confidence that the Census estimates are produced on a sound methodology
- To allow a forum for sharing questions/concerns

## Objectives of the 2011 Census (1)



- **To provide accurate census population estimates**
  - National population estimate is within 0.2% of the truth\*
  - All LA level population estimates within 3% of the truth\*
  - National response rate of at least 94%
  - All LAs have a response rate of at least 80%
- **To provide accurate population characteristics**

\*with a 95% confidence interval

## Objectives of the 2011 Census (2)



- **To provide outputs and delivery mechanisms that meet user needs and ensure confidence in the results**
  - An independent post census assessment of user views on the results, including
    - Quality of the results
    - Timeliness
    - Accessibility and awareness
    - Supporting information, (e.g. metadata)
    - UK coherence



## 2011 Census context



- **The Census is an integrated operation**
  - From address register development to publication of results
- **Coverage assessment and adjustment processes depend on the quality of the previous steps**
  - Overall response rates
  - Variability in response
  - And each of those are dependent on address register quality
- **Every component is designed from the outset with output quality in mind**
  - High quality population estimates in particular
- **It is complex!**

## The 2001 Census



- **Dual System Estimation used for the first time to adjust census results (One Number Census)**
- **Successful in the vast majority of LAs**
- **But localised problems**
  - Adjustments in 15 LAs out of 376, most notably Westminster and Manchester
  - Caused primarily by two issues
    - Localised (i.e. within LA) enumeration failures
    - Out of date planning information

## 4 underlying field issues in 2001



- **Insufficient field staff in areas where most needed**
  - 'One size fits all' approach
    - areas and roles
  - Recruitment challenges
- **Insufficient control in the field**
  - Lack of central management information
  - Insufficient flexibility to respond
- **Information used for planning was out of date**
  - Address register frozen 3 years before census day
  - 'Redeveloped' areas sometimes not identified
- **Local post-back**

## Impact on census estimates



### High variability in response rates

- wider confidence intervals as a result

### Resulting issues with the coverage estimation methodology in some areas

- and insufficient methods for identifying/quantifying bias

## Improvements to field operation – 2011



- Increased follow up resources overall
- More staff where lower response expected
- Assigned staff to a manager, not an area
- Flexibility between areas
- Questionnaire tracking at Household level
- Staff assigned to activities using real-time management information
- Increased community and LA engagement
- Underpinned by the address register

## Improvements to Coverage estimation and quality Assurance methodology

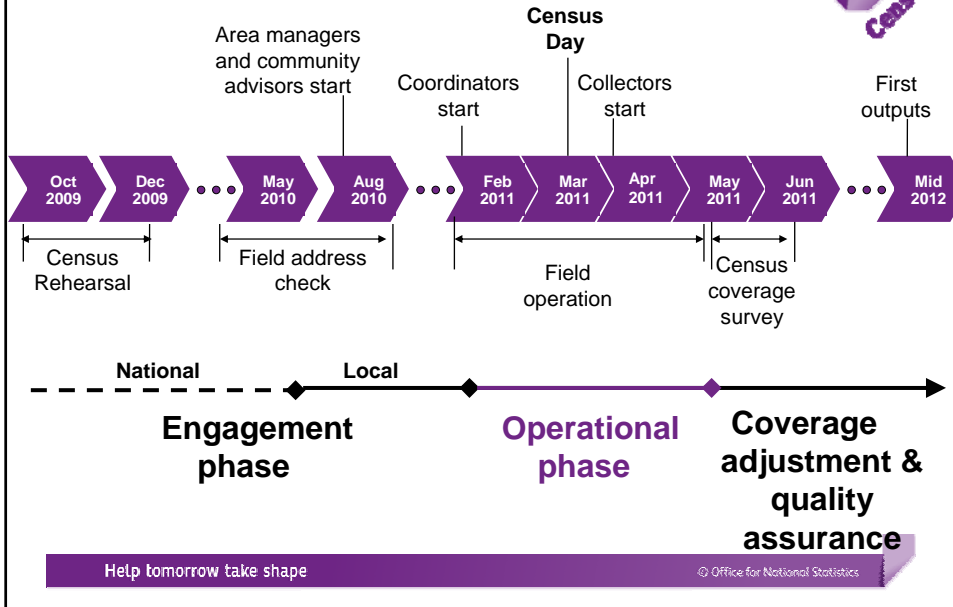


**Methodology has been built on and improved for 2011**

**Improvements since 2001 will be covered at relevant points throughout the day....**

**More detail is available in the paper within your packs:  
'2011 UK Coverage Assessment and Adjustment Methodology'**

## 2011 Census phases



## Coverage assessment and adjustment



Owen Abbott/Paula McLeod



# AGENDA



1. Background
2. Measuring coverage overview
3. CCS
4. Matching
5. Estimation
6. Coverage adjustment (Imputation)
7. Summary

# COVERAGE



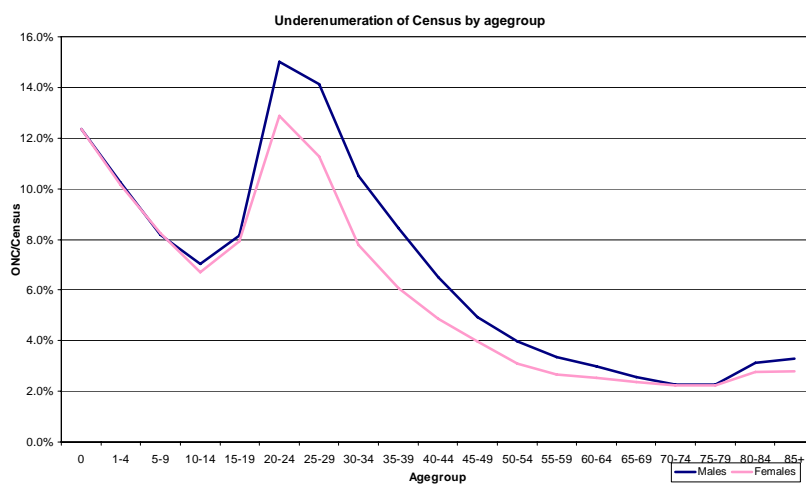
- Some households and persons will be missed by the Census
- Need to adjust the census to take account of this
- Produce estimates by Local Authority and age-sex
- Why?
  - In 2001, ~1.5 million households estimated missed
  - 3.3 million persons (6%) estimated missed (mostly, but not all, from missing 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 (this afternoon)**
  - Checking plausibility of estimates and outputs

# 2001 CENSUS UNDERCOUNT BY AGE-SEX

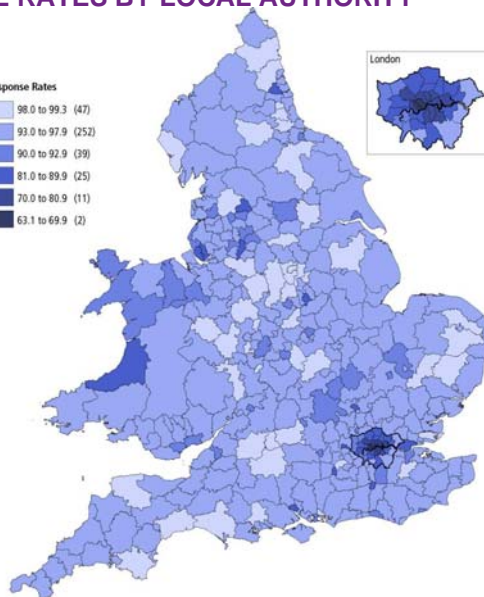


## RESPONSE RATES BY LOCAL AUTHORITY



### Response Rates

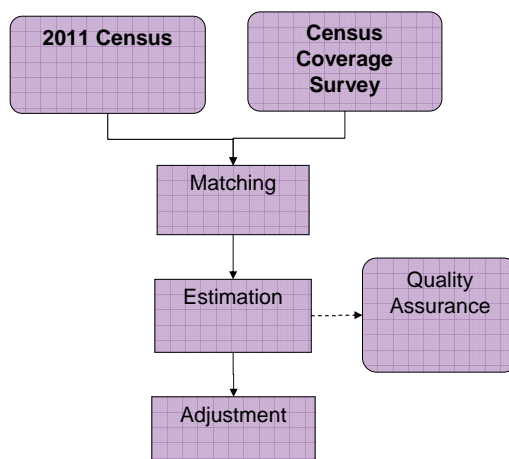
98.0 to 99.3	(47)
93.0 to 97.9	(252)
90.0 to 92.9	(39)
81.0 to 89.9	(25)
70.0 to 80.9	(11)
63.1 to 69.9	(2)



Help tomorrow take shape

© Office for National Statistics

## COVERAGE ASSESSMENT PROCESS OVERVIEW



Help tomorrow take shape

© Office for National Statistics

## WHATS NOT COVERED



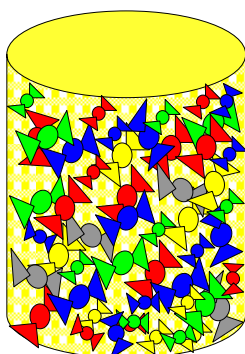
- **Today focused on main parts of the methodology**
- **Things not included:**
  - Overcount
  - Communal Establishments
  - Variance Estimation
  - Full detail of some components
- **These are outlined in the overall methodology paper**

## CAA INTERACTIVE EXERCISE



- **Demonstrating the estimation process**
- **How? By Counting Sweets in a tub**
  - We have a total number of sweets (Population)
  - Different colours (e.g. Sex)
  - We have done a 'census'
  - Those in the census have been marked
- **How do we get from the Census to the population estimate?**
  - Draw a sample
  - Do some estimation

## THE TUB CENSUS



	<b>Marked (Census)</b>
<b>Total</b>	<b>425</b>
Colour 1 (Green)	223
Colour 2 (Purple)	202

- Census count of 425 sweets in the tub (counted sweets have been marked with a sticker)
- Suspect undercount (some sweets don't have a sticker)

### QUESTIONS:

- 1) How many sweets are in the tub?
- 2) How many of each colour are in the tub?

## THE CENSUS COVERAGE SURVEY



- **Key tool for measuring coverage**
- **Features:**
  - Sample of postcodes
    - Measure coverage of households and persons
    - Postcodes cover whole country
  - Large - 330,000 Households
  - 6 weeks after Census Day
    - Fieldwork starting 9th May 2011
  - Voluntary survey

# THE CENSUS COVERAGE SURVEY



- **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 CENSUS COVERAGE SURVEY (CCS)



Adobe Acrobat 7.0  
Document

## THE CCS SAMPLE DESIGN



- **Objective:** design survey to be able to estimate LA coverage
- **Two stage selection:**
  - A) Select 5,500 Output Areas (OAs)
  - B) Select about half the postcodes within the OAs – ‘cluster’
    - Result in selection of clusters of about 60 hhs
- **How are the OAs 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 Lower Super Output Areas (LSOAs)**
- **Uses up to date data sources:**
  - Jobseeker allowance, School census ethnicity, dwelling density, house prices, proportion of 16-29s, crime rate
- **Split into 40%, 40%, 10%, 8%, 2% distribution**
  - Easiest lowest 40%, hardest top 2%
- **Assume OAs have same HtC in LSOA**
- **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 OA per LA/HtC stratum
  - Max 60 OAs per LA (except B'ham and Leeds)
- **Drivers of sample size:**
  - Population size
  - Large undercoverage in 2001
  - Variability in 2001 coverage
  - If HtC patterns changed since 2001

## CCS SAMPLE



- **What does this mean?**
  - Each LA will have its own sample – at least 1 OA for each hard to count level
  - Sample is more skewed to LAs with 'hardest to count' populations
    - Especially London and big cities
  - Sample sizes published

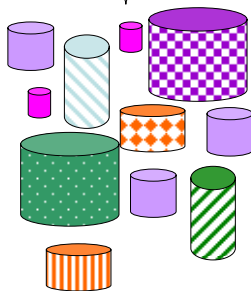
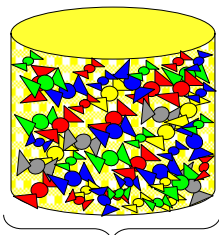




Help tomorrow take shape

© Office for National Statistics

## DRAWING A SAMPLE (CCS)



- **Work in small groups. Each group to collect a cupful of sweets**
  - Variety of sizes of cup
- **Split the cup - going to count most of the cup but some will be missed**
  - Count as many (or as few) as you wish - suggest not all (roughly  $\frac{3}{4}$ ?)
- **Count the numbers of:**
  - 1) sweets in sample that were counted
  - 2) marked and counted sweets in sample
  - 3) total marked sweets

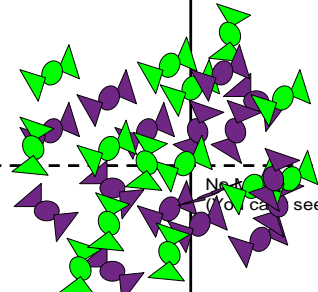
Help tomorrow take shape

© Office for National Statistics

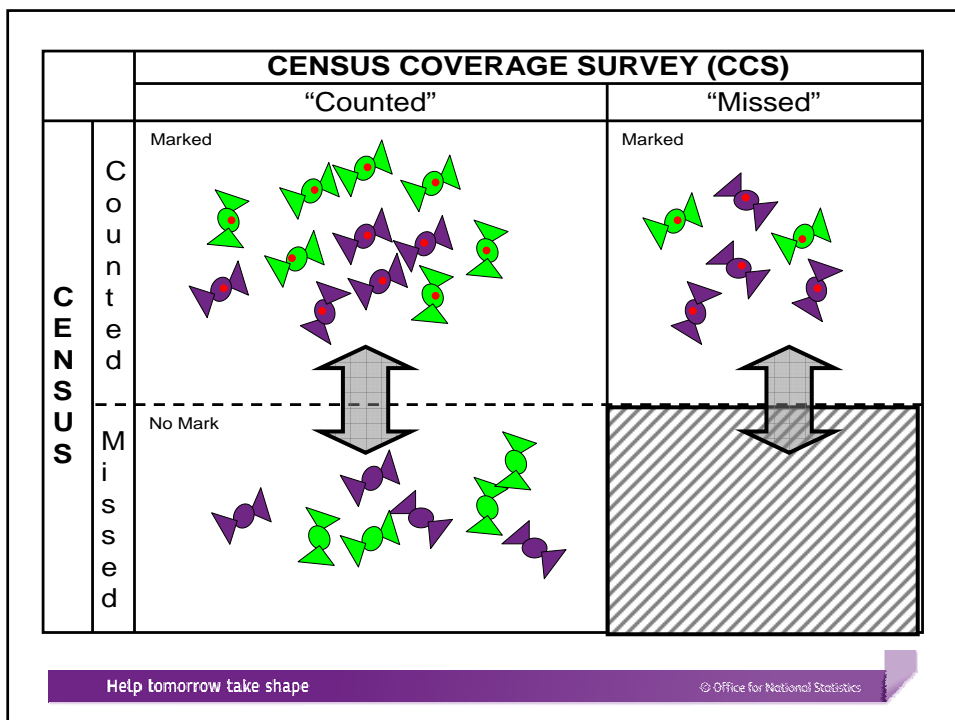
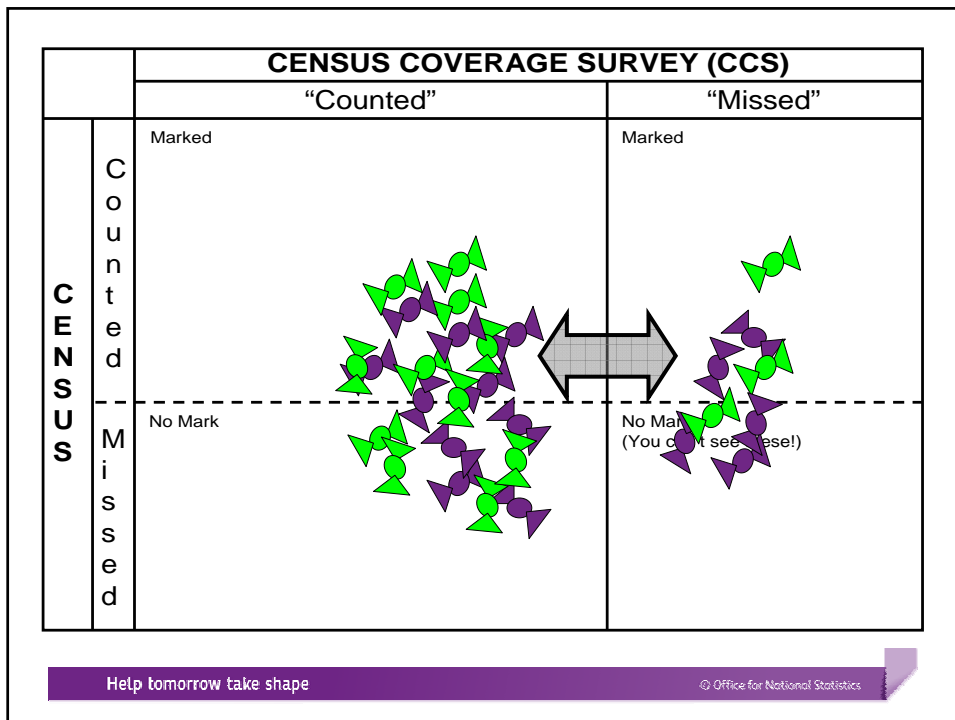
		CENSUS COVERAGE SURVEY (CCS)	
		"Counted"	"Missed"
CENSUS	Counted	Marked	Marked
	Missed	No Mark	No Mark (You can't see these!)

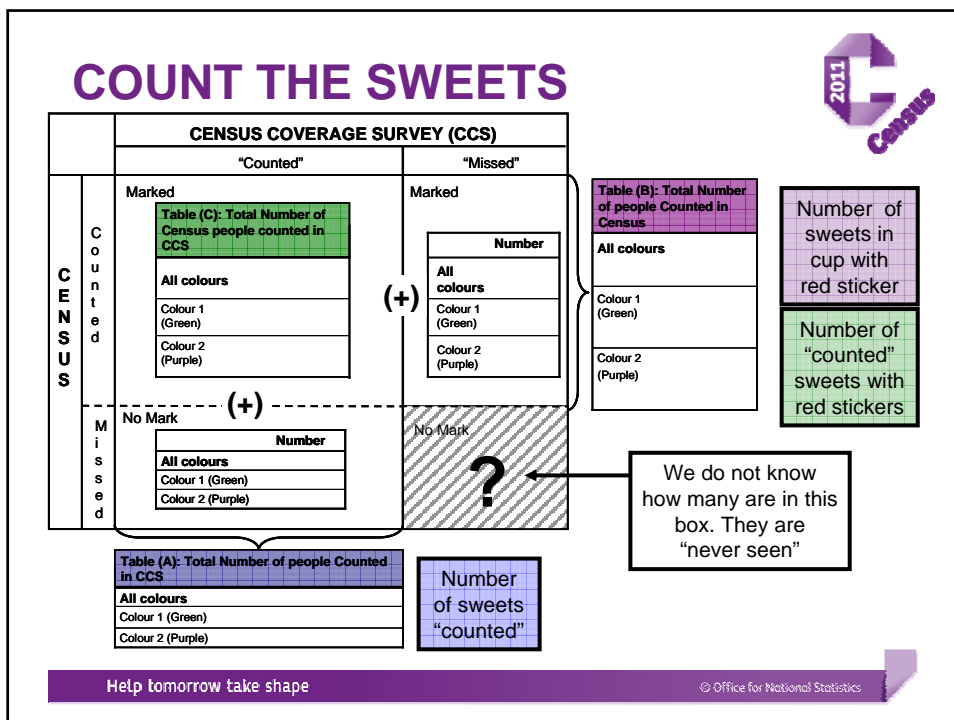
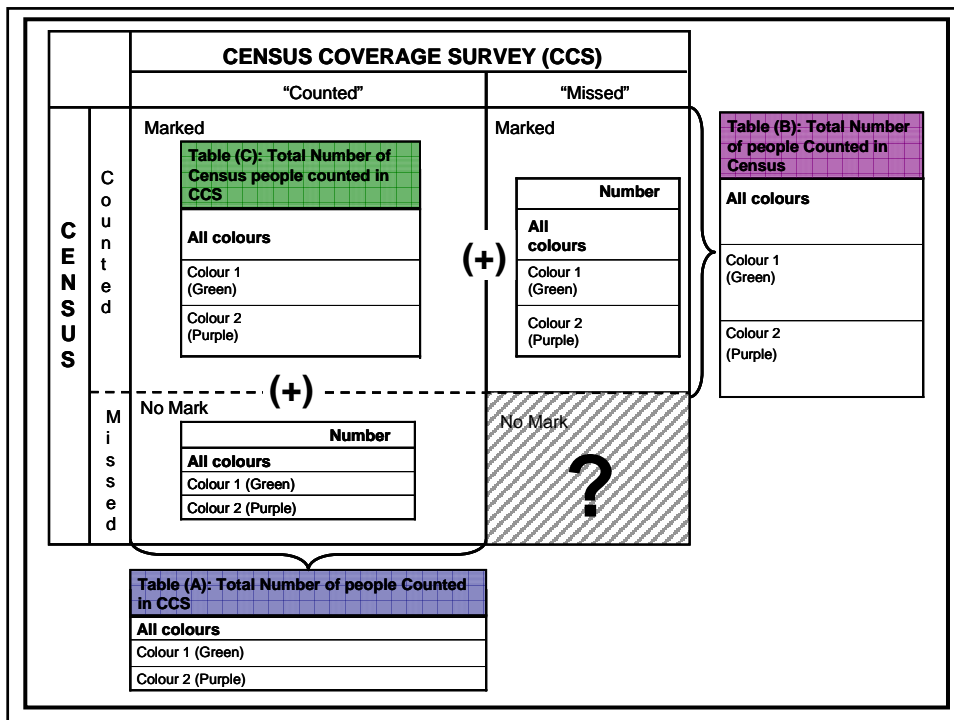
Help tomorrow take shape © Office for National Statistics

		CENSUS COVERAGE SURVEY (CCS)	
		"Counted"	"Missed"
CENSUS	Counted	Marked	Marked
	Missed	No Mark	No Mark (You can't see these!)



Help tomorrow take shape © Office for National Statistics





## END OF FIELD WORK!



Sweets	Representing
<b>Tub</b>	A local authority, the population has been counted in Census and every individual marked with a sticker.
<b>Marked Sweets</b>	Counted in Census
<b>Cup</b>	'Cluster' selected for re-enumeration in the Census Coverage Survey (CCS)
<b>'Counted' from Cup ( 'missed' from Cup)</b>	The people we capture in the CCS (people within postcode cluster not captured in CCS)
<i>Matching output</i> <ul style="list-style-type: none"><li>• <b>Marked Sweets in Cup</b></li><li>• <b>Marked, 'counted'</b></li><li>• <b>Marked, missed</b></li><li>• <b>Unmarked, 'counted'</b></li></ul>	<ul style="list-style-type: none"><li>• Census Count of 'cluster'</li><li>• People captured in both Census and CCS</li><li>• People counted in Census but missed from CCS</li><li>• People counted in CCS missed by Census</li></ul>

## 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)
  - Hierarchical structure – use of surname of head of household
  - 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
  - We expect this to deal with at least 60% of cases
- **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

# MATCHING OVERVIEW



Exact and Probability Matching (automatic)



Clerical Review



Manual Matching



QA

# 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, HoH surname, 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 an expert
  - Expert can defer to a supervisor
- Supervisor must make a decision for all remaining pairs to complete the resolution

## CLERICAL MATCHING



- Matchers attempt to resolve the remaining unmatched records
  - for both CCS and Census records
- Flexible searching at different levels;
  - Postcode and surrounding postcodes
  - Local Authority level
  - Estimation Batch
  - Whole Census
- Either a match is found, or the supervisor confirms that a record is unmatchable (with a reason where appropriate)
- Process completes when all CCS and Census (in CCS postcode) records have a match, or an unmatchable status



# 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	JACK ANTMONY DONEGAN	18041992	2	PHILIP DONEGAN	1111988
3	CHLOE MARIE DONEGAH	missing	3	JACK DONEGAN	18041992
			4	CHLOE DONEGAN	6011995

## DATA AFTER MATCHING



- We have for the sampled areas (about 5,500 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

# POPULATION COUNTS



**We have two counts of the numbers of sweets in your cup/postcode cluster:**

## Count 1- the Census

- *We can see unmarked sweets so we know the Census count is lower than the "truth"*

## Count 2- the Census Coverage Survey

- *A second count of a small sample of the Census*
- *Counted some "extra" sweets in our sample (unmarked)*
- *We can still see some unmarked, uncounted sweets. Still haven't reached the "truth"*

**Can we estimate the number of uncounted sweets and so improve our population estimate?**

# 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 EA 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



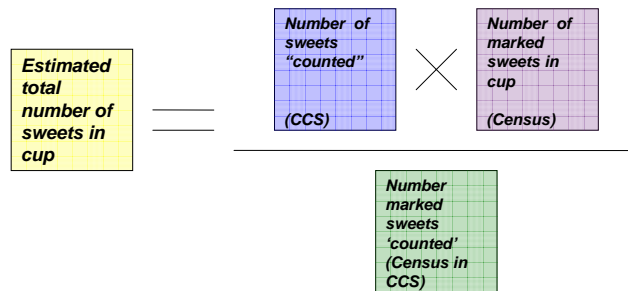
- DSE estimates adjustment for those missed in both Census and CCS in each cluster by age-sex group

		Counted By CCS		
		Yes	No	TOTAL
Counted By Census	Yes	$n_{11}$	$n_{10}$	$n_{1+}$
	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

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

# ESTIMATING POPULATION OF CUP



Help tomorrow take shape

© Office for National Statistics

## • Counting the number of sweets in the cup

Total Number of Sweets in Cup:

i) Census Count (Total number of marked sweets) =

ii) "Union" Count (Census plus number of "new" sweets found in CCS) =

## • Can we do better? How many sweets are "never seen"?

Dual System Estimation (DSE) of Total Number of Sweets in Cup

(calculating the total number of sweets in cup by including an estimate of the number of sweets **MISSED** from **BOTH** Census and CCS)

$$\text{Total Number of Sweets in Cup (i.e. DSE)} = \frac{\text{Count}(\text{CCS}) \times \text{Count}(\text{Census})}{\text{Count both}(\text{CCS}\&\text{Census})} = \frac{\text{A} \times \text{B}}{\text{C}} = \text{D}$$

N.B. Do not round off your figures at this stage.

(see presentation and/or workbook for a full-explanation of how we arrive at above equation for the DSE)

## ESTIMATING THE CUP POPULATION



Excel Workbook: Calculating  
DSE for Interactive Exercise.



Excel Sheet: CUP TOTAL

## 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



- **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 35 or more in total
  - 36 LAs have enough sample to be EAs themselves
  - EAs are formed from contiguous LAs
    - Respecting Welsh border
  - Have now published EAs

# ESTIMATION AREAS



ESTIMATION  
AREAS

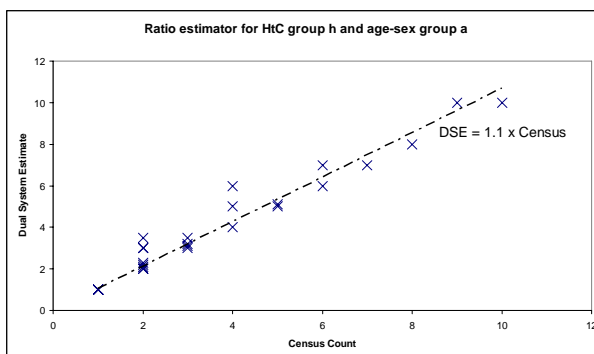
Help tomorrow take shape

© Office for National Statistics

# RATIO ESTIMATION



- **Step 1 – the 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")



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.

Help tomorrow take shape

© Office for National Statistics



## RATIO ESTIMATION



- Step 2 – assume the relationship holds across the non-sampled areas and predict using relationship
  - Apply the adjustment factor to the total census count for an Estimation Area
  - (ratio calculated and applied for groups distinguished by geography, HtC and age-sex group)
  - Ratio estimator is:

$$\hat{T} = \frac{\sum_{sample} \frac{n_{1+} n_{+1}}{n_{11}}}{\sum_{sample} census} \sum_{all} census$$

## RATIO ESTIMATION



- Ratio of census counted to estimated population gives the proportion of under-enumeration in the Census (adjustment factor)


$$\text{adjustment factor} = \frac{\text{Best estimate of population of the cup (DSE)}}{\text{Number of marked sweets in cup}}$$


- Use this adjustment factor to calculate the number of sweets in the tub

$$\text{Number of sweets in tub} = \text{adjustment factor} \times \text{Number of marked sweets in tub}$$

• **Number of Sweets in the Tub.**

1. Calculate Adjustment factor: Adjustment factor =  $\frac{\text{Estimate of sweets in cup}}{\text{Census count of cup}}$
2. Assume the ratio of counted to uncounted sweets in the cup is that same as that in the tub.
3. Calculate the number of sweets in the tub by multiplying the Census count of the tub by the coverage weight calculated for the cup.

**Total Sweets in Tub** =  $\frac{\text{DSE of cup}}{\text{Census count of Cup}}$  x Census count of tub =  x 425 =

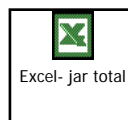


*(This is your groups single estimate of the total number of sweets in the tub. We do not take a single estimate in isolation- we will use the estimates from all groups to provide a robust estimate of the tub total, along with a measure of variability in this estimate.)*

## ESTIMATING POPULATION OF TUB



Excel Workbook: Calculating DSE for Interactive Exercise.



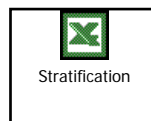
Excel Sheet: TUB TOTAL

## ESTIMATING POPULATION OF TUB



- Estimated population for the cup and the tub based using total counts
- We can do better than this by stratifying (doing the two colours separately)
- Repeat the analysis for the separate colours

Excel Workbook: Calculating  
DSE for Interactive Exercise.



Excel Sheet: STRATIFICATION

## DISCUSSION



- **How did we do?**
  - Several different estimates of adjustment factor so we have several population estimates.
- **How does the range of estimates compare to “truth”?**
  - Are they all better than the census count?
- **How good are the estimates of each sweet type?**
  - Expect that as we consider smaller populations with poorer Census (or CCS) coverage that estimates become more variable

## 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 ('cup') 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 ('cup')
  - The estimate will come from the address register and supplementary information

## 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)
- **We place households into these dummies using a best fit approach**
  - E.g. use try to use same accommodation type and 'copy' records from nearby

## SUMMARY



- Taken through end to end coverage process
  - CCS
  - Matching
  - Estimation
  - Adjustment
- Shown how estimates are calculated
  - Not included all the detail
  - Semi-realistic example
  - Demonstrated the key principles
- Next stage is the QA of those estimates

## Questions?

[owen.abbott@ons.gsi.gov.uk](mailto:owen.abbott@ons.gsi.gov.uk)





# Overview of Quality Assurance (QA)

Jonathan Wroth-Smith

## Objectives of the 2011 QA



### Accuracy

- Ensure 2011 Census outputs are fit for purpose
- Understand differences between Census and rolled-forward mid year estimates
- Ensure Census population characteristics are accurate

### Transparency

- Methods and sources
- Decision making process
- Stakeholder liaison
- Contingency and when to apply
- Defining quality measures to publish with results

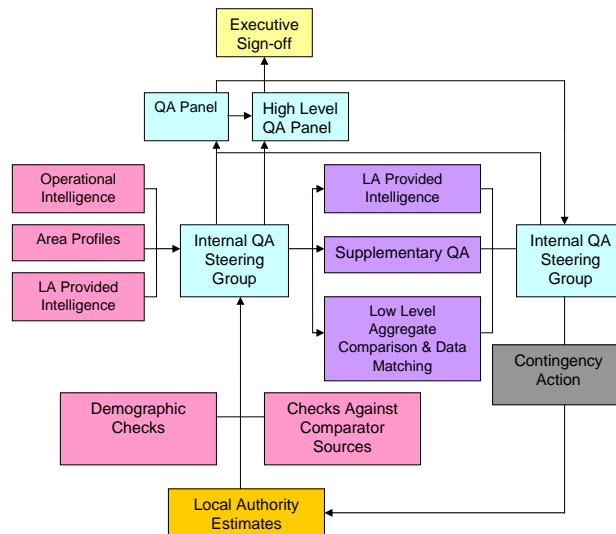


## Accuracy is measured through QA checks

- **Early Extract checks**
- **Core QA for LAs**
  - Comparator checks
  - Demographic analysis
  - Operational Intelligence
  - Area Profile
- **Supplementary QA**
  - Drilldown to lower geographies
  - More detailed checks on population sub-groups
  - Cross checks to look back at changes after different processing stages
- **Checks for higher geographies e.g. regional/national**
  - Cumulative checks



## Overview of the 2011 Census QA Process





## Lessons learned from the 2001 QA process



- **More information about census data collection would have helped**
- **Actions should be in place to act upon when census estimates were implausible**
- **More use should be made of local data and knowledge**
- **Needs of customers should be incorporated where possible in the development of QA process**
- **Data quality monitoring should be targeted at errors that could have a substantial impact on outputs**

## Early Extract data



- **All census records scanned on a daily basis delivered directly from the data processing site**
- **Starts 2 weeks before census day and continues for 42 weeks**
- **Helps build an overall picture of any quality issues in the census data**
  - **opportunity to identify systematic quality issues and take corrective action**



# Interactive group exercise

Beth Moon



## Quality Assurance group exercise

**Aim of the exercise is:**

**- to provide an understanding of the range of sources available for quality assurance and how they are used**

**You will be shown information for two fictitious local authorities**

- **Scenario A (Candytown)**
- **Scenario B (Sweet City)**

## QA group exercise – how it will work (1)



**You will be presented with 5 sources of information one at a time**

- Comparator analysis
- Cumulative checks
- Demographic analysis
- Operational Intelligence
- Area profile

For each source:

1. Introduction to QA source – what it is, how we intend to use it
2. Example for scenario A (Candytown) of what the information would look like
3. Scenario B (Sweet City) is for you to work through
4. Follow up of what you found and more information about the source

## QA group exercise – how it will work (2)



Your role as a group is to quality assure the census population estimate for Sweet City

After each piece of information is presented you should consider:

- **What does the information tell you about the census population estimate for Sweet City?**
- **If you were responsible for signing off the estimate for this area, would you be happy to do so? If not, why not?**
- **How does the information provided relate to what you have already learned about the area?**
- **Is there any more information you would like to see and why?**
- **What conclusions would you draw?**

Please nominate one person from each group to write a summary of your comments for each section in the space provided on the worksheet



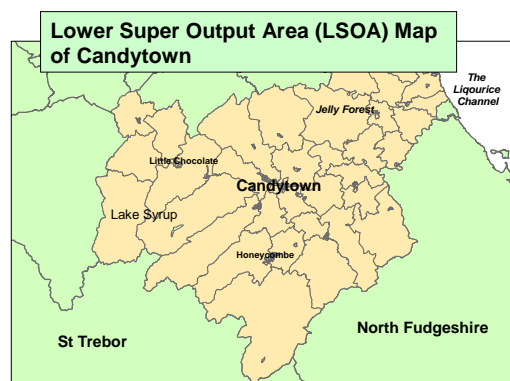
## Setting the scene... Scenario A: Candytown

Brief area profile

- Located in a predominantly rural area of Northern England
- 2001 Census population count – 102,000
- Low population growth as shown by trend in Mid Year Population Estimates

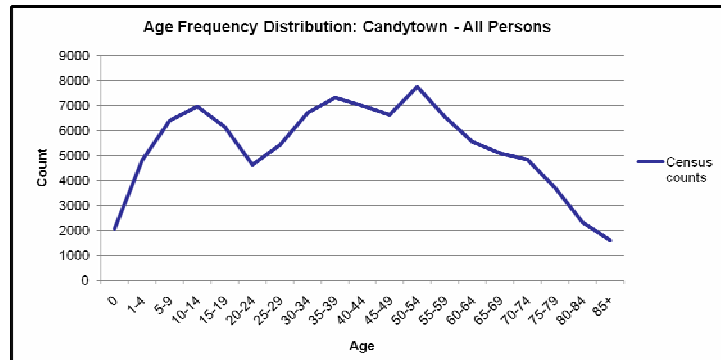


## Scenario A – Candytown





## Scenario A - Candytown

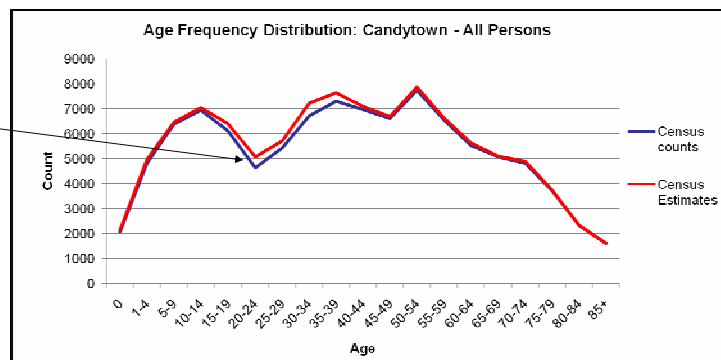


Blue line = Population counts based on raw census data



## Scenario A - Candytown

Dip in young people who left the area for university



Blue line = Population counts based on raw census data

Red line = Data after coverage estimation i.e. once those people who were not included on a census questionnaire have been estimated



# Comparator analysis



## Comparator analysis

**Census estimates will be validated by comparing them to a range of comparator data sources, including:**

- Administrative data: such as School Census and Patient Register
- Survey data: such as the Integrated Household Survey

**A series of checks have been developed, using the information available in the different comparator sources e.g.**

- Age and sex check: comparing age distribution of Census estimates to a range of sources
- Students check: the age sex distribution of students against Higher Education Statistical Agency data to validate student estimates

**The checks will be run at different levels of geography but LA is the default level**

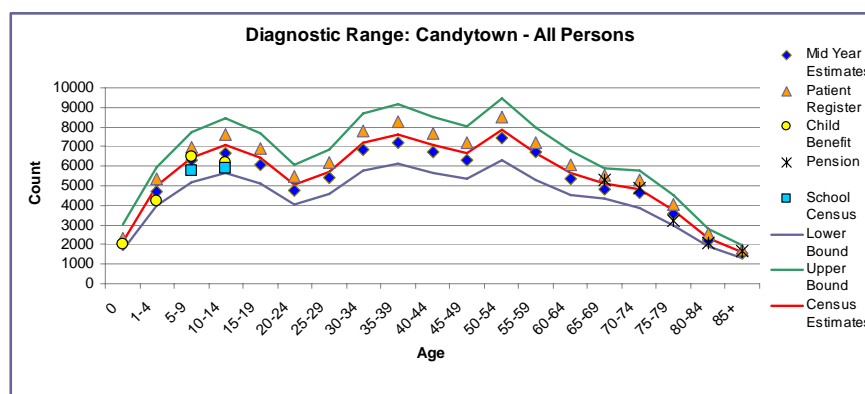


## Comparator analysis

- Comparators and Census will not match exactly due to:
  - Definitional differences- e.g. School Census does not include independent school children
  - Coverage reasons- e.g. Patient Register is known to over/under count people
- Known differences between the comparators and Census will be taken into account to create tolerances (upper and lower bounds) within which Census estimates might be expected to fall
- A Census estimate that fell outside of these bounds would require further investigation



## Comparator analysis Scenario A - Candytown





## Your turn!

Look at the information presented for scenario B – Sweet City

For the next few minutes.....

- Look at the comparator analysis information and consider the following questions;
  - **What does the information tell you about the census population estimate for Sweet City?**
  - **If you were responsible for signing off the estimate for this area, would you be happy to do so now? If not, why not?**
  - **How does the information provided relate to what you have already learned about the area?**
  - **Is there any more information you would like to see and why?**



## Further information on comparator checks

- We showed you an example of an age check against a diagnostic range
- The comparator analysis revealed that the Census estimates are below the lower bounds for: under 1s, 1 to 4 yr olds, 25 to 29 yr olds and 30 to 34 yr olds
- There are a number of key checks available to the QA team that use a variety of comparator data sources. Some checks are for individuals and others are at the household level
- All Local Authority level estimates run through automated QA checks



## Key automated comparator checks and data sources



QA Check	Comparator dataset
Age and sex	<ul style="list-style-type: none"> <li>•Patient Register</li> <li>•Mid-year Population Estimates</li> <li>•School Census</li> <li>•Child benefit/pensions data</li> </ul>
Household Number and Average Size	<ul style="list-style-type: none"> <li>•Council Tax</li> <li>•Address Register</li> <li>•Patient Register</li> <li>•Communities and Local Government household projections</li> </ul>
Ethnicity	<ul style="list-style-type: none"> <li>•Population Estimates by Ethnic Group</li> <li>•Integrated Household Survey</li> <li>•School Census</li> <li>•Independent Schools data</li> </ul>

## Key comparator checks and data sources



QA Check	Comparator dataset
Students	<ul style="list-style-type: none"> <li>Higher Education Statistics Agency (HESA)</li> <li>Further Education Student Numbers from Business, Innovation and Skills</li> </ul>
Armed Forces (Home/Foreign)	<ul style="list-style-type: none"> <li>Defence Analysis Statistics Agency</li> <li>US Armed Forces</li> </ul>
Migration (internal)	<ul style="list-style-type: none"> <li>Patient Register</li> </ul>
Migration (international)	<ul style="list-style-type: none"> <li>Patient Register</li> <li>International Passenger Survey</li> <li>Migrant Workers Scan</li> </ul>



## Calculating bounds

Bounds are the range of values within which the census estimate would be considered plausible.

These values are crucial in identifying areas that require further investigation.

Two main approaches

1. Diagnostic range approach

- used when there are two or more comparators
- ranges are calculated based on the variation between the sources

2. Quality assessment approach

- method used when there is only one comparator source
- based on quantifying known quality issues with the comparator



## Cumulative checks

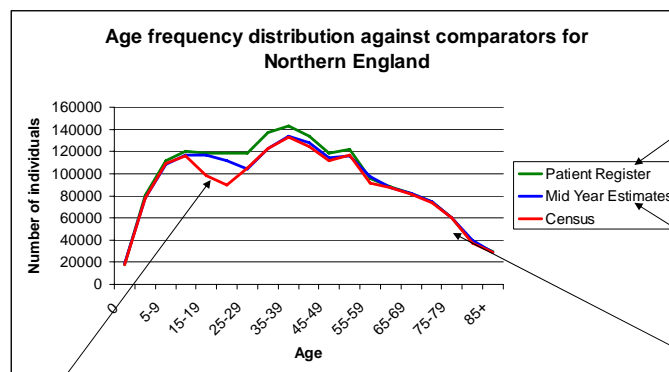


## Cumulative checks

- Works like a comparator check but allows us to aggregate data and perform checks at a higher level
- This enables the QA team to:
  - pick up issues that may not be apparent at a low level, but can be seen at a higher level
  - see if issues identified at a lower level e.g. LA are localised, or apply to a wider area e.g. regionally
- Cumulative checks may help to identify systematic bias in our estimation
- The level at which data is aggregated up by can be customised, for example we could look at only those LAs with universities



## Cumulative checks - Scenario A



At this level there are no bounds, only direct comparisons against comparator data sources

Census estimates look very similar to comparators for older age groups

Census estimates are lower than both comparators for ages 15-25



## Your turn again!

Look at the information presented for South East England

For the next few minutes.....

Look at the cumulative checks information and consider what the information is telling you...

- **If you were responsible for signing off the estimate for Sweet City, would you be happy to do so now? If not, why not?**
- **How does the information provided relate to what you have already learned about the area?**
- **Is there any more information you would like to see and why?**



## Further information on cumulative checks

- We showed you an example of a cumulative age check using comparator data sources
  
- The check did not indicate a problem at any age group for the grouping of local authorities in the region of South East England
  
- Cumulative checks can be conducted:
  - **For any check e.g. ethnicity check, student check**
  - **At standard geographies or bespoke geographies**
  - **At any stage of processing**



# Demographic Analysis

## Demographic analysis



- **A number of demographic indicators will be used to provide further validation of the Census estimates**
  - Sex ratios
  - Fertility rates
  - Mortality rates
- **More demographic checks will be conducted than were used in the 2001 QA process**

## Demographic analysis



**Demographic analysis is a key part of the Quality Assurance process as it:**

- Is based on accurate and timely registration data
- Provides a more comprehensive view of the Census estimates by highlighting issues that may not be identified in a direct count comparison
- Will further inform the expert demographers who will be assessing the accuracy of the Census estimates

## Sex ratio

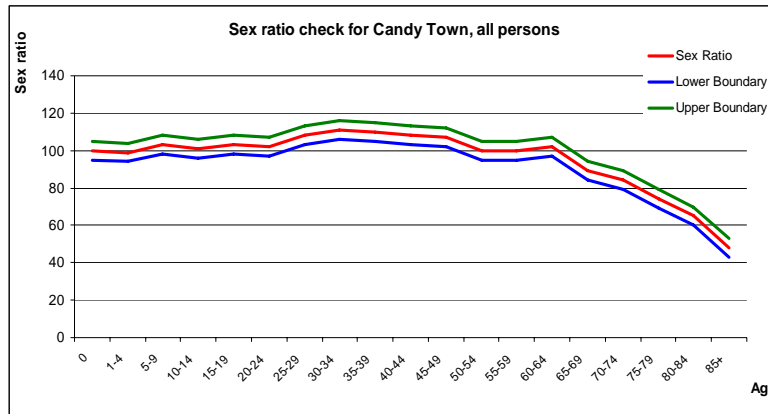


Sex ratio is measured as the number of males in a population per number of females, expressed per 100 females

For example, a sex ratio of

- 80 means that there are only 80 men for every 100 women
  - shortage of men
- 100 means that there is a perfect match in number between men and women
- of 110 means that there are only 100 women for every 110 men
  - shortage of women

## Demographic analysis Scenario A - Candytown



Help tomorrow take shape

© Office for National Statistics

## Your turn again!



Look at the Demographic analysis presented for Sweet City

For the next few minutes.....

- Consider what the information is telling you...
  - If you were responsible for signing off the estimate for Sweet City, would you be happy to do so now? If not, why not?
  - How does the information provided relate to what you have already learned about the area?
  - Is there any more information you would like to see and why?

Help tomorrow take shape

© Office for National Statistics



## Further information on Demographic Analysis

We showed you an example of a demographic check – a sex ratio

The demographic analysis did not indicate a particular issue relating to either men or women

We also conduct other demographic checks such as

- fertility rate
- mortality rate

These checks might detect problems that are not seen by the other indicators



# Operational Intelligence





## Operational Intelligence

- Information on the field operation and data processing diagnostics will be made available for quality assurance purposes
- These sources will provide a useful early indication of data quality issues to consider alongside the other sources
- Field information includes:
  - Key information on return rates
  - Notes from field staff debriefs
  - Information on field incidents
  - Key information from the Questionnaire tracking system
- Data processing diagnostics includes:
  - Key information from reports on different stages of data processing



## Operational Intelligence Scenario A – Candytown

### Field information summary

<b>Forms posted:</b>	<b>62,514</b>
<b>Forms undelivered/deactivated:</b>	<b>31 (0.05%)</b>
<b>New addresses:</b>	<b>38 (0.06%)</b>
<b>Number of Dummy forms completed:</b>	<b>2050</b>
% absent households:	4%
% refusals:	4%
% non-returns:	82%
% holiday homes:	4%
% second residences:	3%
% vacant:	3%
<b>Final return rate:</b>	<b>97%</b>



## Operational Intelligence Scenario A – Candytown

### Census Field information

#### Field Incidents

Flooding in the area around Lake Syrup resulted in access problems for some collectors

#### Field staff debriefs

Collector debrief

- For a short period of time, adverse weather conditions and flooding caused some difficulties with making follow up attempts



## Operational Intelligence Scenario A – Candytown

### Coverage estimation diagnostics and Census Coverage Survey (CCS) field information

#### Estimation Area detail:

The local authorities included in this Estimation Area are:

Candytown, St Trebor, North Fudgeshire, Puddingmouth, and Mid Sherbertshire.

Number of CCS postcodes:	180
Number of households listed:	3,117
% CCS responses:	92% (expected 80-90%)
% refusals:	2.1% (expected 5-15%)
% non contact/other:	5.9%

The overall coverage for this estimation area is 97%



## Your turn again!

Look at the information presented for Sweet City

For the next few minutes.....

Look at the operational intelligence provided and consider what the information is telling you...

- **If you were responsible for signing off the estimate for Sweet City, would you be happy to do so now? If not, why not?**
- **How does the information provided relate to what you have already learned about the area?**
- **Is there any more information you would like to see and why?**



## Further information on Operational Intelligence

- We showed you some examples of the type of information available to the QA team on the census operation.
- Census field information revealed:
  - **lower coverage in a few MSOAS in the East of the LA**
  - **some enumeration issues**
  - **CCS field operation was reasonably successful**

Other information available may include:

- **Detailed breakdown of return rates for different geographies**
- **Additional field information e.g. from follow up worksheets**
- **Processing diagnostics from other processing stages**



# Area Profile



## Area Profile

ONS will create an area profile containing information for each local authority

It includes:

- Statistical information on the LA (e.g. Demographic, economic, housing)
- Information on Communal Establishments (Location, type, size, etc)
- Other intelligence about the area (e.g. major building developments)



## Area Profile Scenario A - Candytown

### Statistical Information

Population:

2001	2002	2003	2004	2005	2006	2007	2008	2009
102,000	101,800	101,900	102,200	102,500	102,500	102,700	102,600	102,700

Source: *Mid-year Estimates (ONS)*

Language Profile (First Language):

5-9	10-14
99% English	99% English
1% Other	1% Other

Source: *School Census*

Ethnicity

97% White
1% Asian
1% Mixed
1% Other

Source: *Mid-year Estimates by ethnicity (ONS)*

Very little population change

Suggests that there are not many migrants and/or people who may have difficulties with the census form



## Area Profile Scenario A – Candytown Communal establishments (top 5)

Type	No. of usual residents	OA Code	Postcode	Further info
Care home	90	00CAND1123	CY15 0XX	Elderly Care
Care home	80	00CAND1232	CY15 9YI	Elderly Care
Hospital	60	00CANE0112	CY12 6TR	Local general hospital
Travel and Leisure	20	00CANF1021	CY12 8AD	Caravan site
School	15	00CANH7638	NF13 7RR	Independent boarding school



## Area Profile Scenario A - Candytown

### Other Intelligence

No significant building of new homes in the last 15 years (local development)

Development plans underway for a large new caravan park in the area (News story)



## Your turn again - for the last time!

Look at the information presented for Sweet City

For the next few minutes.....

•Look at the area profile provided and consider what the information is telling you...

- If you were responsible for signing off the estimate for Sweet City, would you be happy to do so now? If not, why not?
- How does the information provided relate to what you have already learned about the area?
- Is there any more information you would like to see and why?

## Further information on Area Profile



- We showed you some of the types of information that may appear in the area profile
- The area profile revealed a fairly high proportion of children who didn't have English as a first language
  - lots of speakers of Eastern European languages suggests a large amount of migrants
- Other information may include:
  - **Correspondence with LAs on population estimates since the last census**
  - **Information provided by LAs through the Census Local Partnership Plans**
  - **Further trend data and other evidence about the area**

## So what are the conclusions for Sweet City?



**We have underestimated young migrant families who moved to Sweet City to take advantage of the new jobs and housing associated with the Sweet factory. It is a very localised issue that wasn't fully accounted for in the coverage estimation process.**

## Key messages



What did you learn from this exercise?

- The core package of material for quality assuring census estimates includes a wide range of information
- Information is considered as a whole to provide the full picture
- The purpose of setting bounds around comparator sources is to provide an indication of areas that may require further investigation
- In some cases e.g. Candytown enumeration is successful, the amount of estimation required is minimal and signing off the estimates is straightforward
- In other cases e.g. Sweet City the quality assurance is more complex and further work may be required

## Questions?







# Supplementary Quality Assurance

## Supplementary QA



### What is supplementary QA?

- The QA conducted when issues that have been identified or aspects of the data are difficult to explain

### What will it include?

- Additional QA checks
- Drilldown and cross checks
- LA supplied evidence
- Low level aggregate comparisons and matching

## Supplementary QA checks



- **Additional comparator checks available when a potential issue is found**  
e.g school census data on first language used to identify groups not identified though the ethnicity check
- **Some additional checks will be pre-specified others will be carried out during live operations**
- **Not all supplementary checks will be conducted for all areas**

## Drilldown and cross checks



- **Carried out in LAs where census estimates fall outside tolerances**
- **Drilldown – look at lower geographic levels**
- **Crosschecks – information varies by processing stage but will give an indication if a process has created an unexplainable change**

## LA provided intelligence



### **LAs have provided a range of information during build up to the Census**

- QA studies - Information/analysis provided by LAs e.g. reports into population sub groups, accuracy of admin data
- Census Local Partnership Plans
- Communication with ONS on population estimates since last census

**This information will be used to get the best possible understanding of an LAs population in advance of the census**

## Low level aggregate comparisons



**Access to administrative data at record level allows comparisons to be made at low geographic levels**

**It is also possible to match at record level between the census and other sources**

### **Used to:**

- Validate coverage estimation using admin data
- Validate within household count



# Process of getting to a final estimate

## Process of getting to a final estimate



### Supplementary QA

Issue resolved following further investigation

Issue remains, Action required

- What happens next?
- Revisit coverage estimation
- Explore local and national contingency options
- Implement contingency
- Redo QA
- Sign-off estimates

## Revisit coverage estimation



**A number of options are available for revisiting coverage estimation in order to make an adjustment for issues identified**

**For example,**

- Changing estimation areas
- Changing characteristics by which DSE is stratified

## Local contingency options



- **Adjust on the basis of other census information collected**
  - Visitor information
  - Second residences
- **Adjust by calibrating to external sources**
  - Geographic areas
  - Population sub-groups

# National contingency options

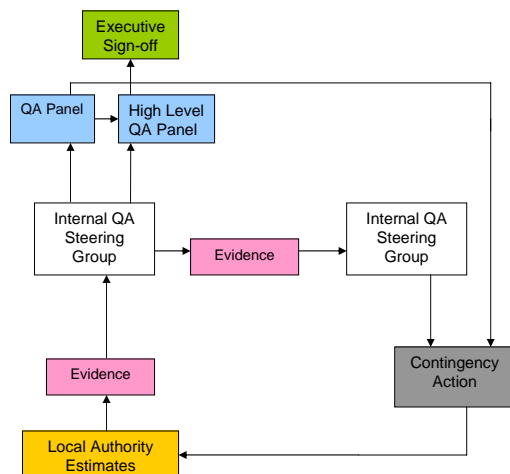


Cumulative checks will show the emerging estimate for England & Wales

## Examples of adjustments that may be made:

- Longitudinal Study adjustment
- Overcount adjustment
- Adjust by calibrating to external sources

# Sign-off process



## Sign-off process



### Groups in sign-off:

- Internal QA steering group
  - Main QA Panel
  - High level panel
- 
- **Executive sign off prior to publication**
- 
- **Approach in 2001 used a single main QA panel**

## Internal QA steering group



- **Working level group with experts from Census, Methodology and the ONS Centre for Demography**
- **Meeting on a day-to-day basis**

### Objectives:

- To focus on difficult LAs and provide a steer supplementary QA analysis
- To advise on possible contingency options
- To reduce the burden on the main QA panel

## Main QA Panel



- Panel consisting of representatives from different areas within ONS and the Welsh Assembly Government
- Meeting weekly

### Objectives:

- To review census estimates for every local authority
- To recommend acceptance or rejection
- To identify supplementary QA analysis

## High level panel



- Representatives from across ONS, the other UK statistical agencies and independent external experts
- Meeting every six to eight weeks

### Objectives:

- Review the emerging regional/national picture
- Advise on necessary regional/national adjustment
- Review supplementary QA and contingency action taken for local area estimates





# Wrap up

## Coverage Estimation Improvements Since 2001



- **Variability managed in the field**
- **Residual variability explicitly understood through address register and questionnaire tracking data**
- **Bias assessment for every LA:**
  - Address register / questionnaire tracking data
  - Linkage to ONS survey data
- **Default position:**
  - We understand sources and issues
  - Tend towards adjustment if significant

## Quality Assurance Improvements Since 2001



- **Early Extract**
- **Work carried out with users to understand local sources**
- **Greater use of demographic analysis**
- **Improvements to checks carried out**
  - Wider range of checks
  - Improvements in quality of comparator data
  - Greater understanding of comparator sources
- **Longitudinal Study analysis**

## Quality Assurance Improvements Since 2001



- **Use of cumulative analysis to understand regional/national issues**
- **Greater use of 'drill down' to consider below LA level discrepancies**
- **Collecting information on short-term migrants and second residences**
- **Use of administrative data sources at very small geographies**

## Summary



- **Doing a good Census and CCS is really important**
- **Inputs to coverage assessment process improved**
  - And far better understood
- **Bias assessment methods matured significantly**
  - Tend towards adjustment if significant
- **QA methods matured significantly**
  - Tend towards adjustment if all point in the same direction
  
- **Major advances since 2001**

## Questions?





**Thank you**

**Please complete your delegate  
feedback form  
We hope you have a safe journey  
home**



**Reference slides**

## Diagnostic range calculations



The DR is calculated as follows:

1. **Max(comparator) = X**
2. **Min(comparator) = Y**
3. **Range (R) = X – Y**
4. **Midpoint (M) = (X+Y)/2**
5. **Diagnostic boundaries**
  - Upper bound (UB) = M+R
  - Lower bound (LB) = M-R
6. **Constrain the UB and LB to limits to make them more plausible**
7. **Diagnostic range (DR) = UB-LB**