Office for National Statistics

Statistical Disclosure Control in the 2011 UK Census: Swapping Certainty for Safety

Joe Frend

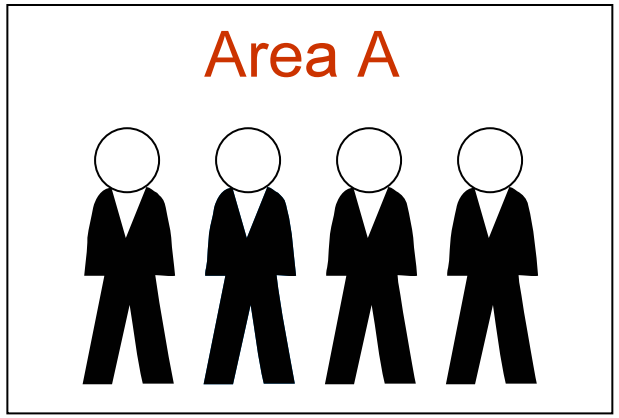
Statistical Disclosure Control, Office for National Statistics

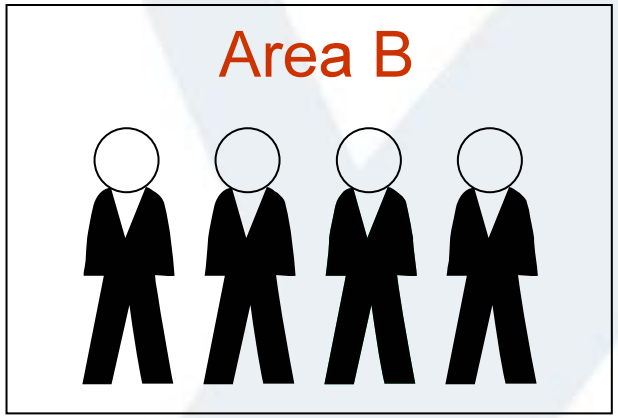
Overview

- Introduction and definitions
- UK SDC Policy
- Household methodology
- Communal establishment methodology
- Sparsity and threshold rules
- Doubt measure
- Summary
- Q&A

Record Swapping

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





Geography

≈ 100 Delivery Groups (DGs) in England & Wales

(≈ 500,000 persons & ≈ 200,000 households per DG)

- 1 - 7 Local Authority Districts (LADs) in a DG
- ≈ 25 Middle Super Output Areas (MSOAs) in an LAD

  (≈ 7,500 persons & ≈ 3,000 households per MSOA)
- ≈ 20 Output Areas (OAs) in an MSOA

  (≈ 300 persons & ≈ 120 households per OA)

Disclaimer

We will not reveal specifics of the record swapping methodology such as:

- Swap rates
- Risk variables
- Matching variables

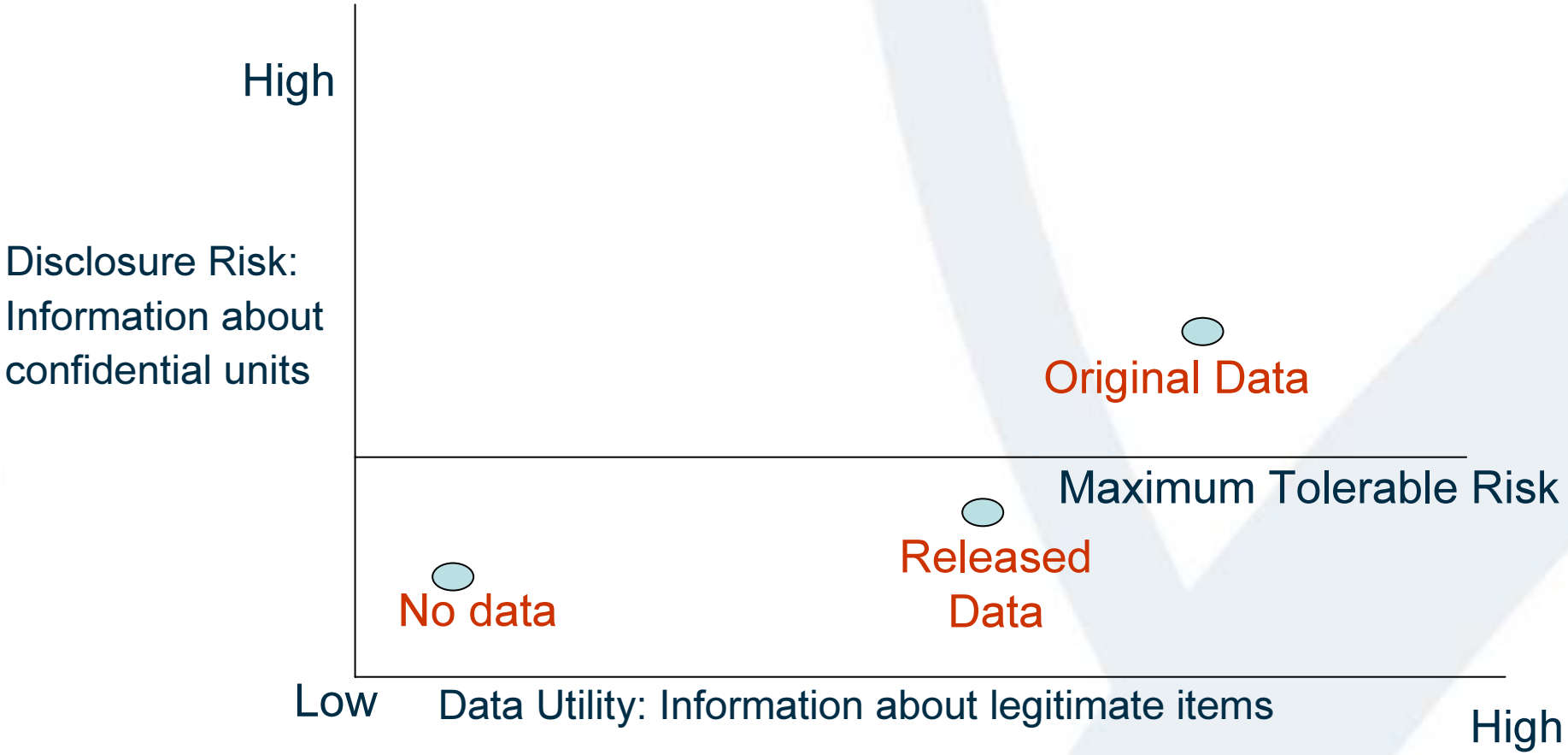
Definitions

- Communal Establishments (CEs): An establishment providing managed residential accommodation

- CE Type: The broadened category in which a CE will appear as in the Census output tables

- Residents: All persons living in a CE
  - Client: Non-staff residents that the CE caters for *(eg. patients of a hospital, clients of a hotel etc.)*
  - Staff: Staff / Owners living in a CE
  - Family: Family members / partners that live in a CE with either a member of staff or a client

Risk – Utility balance



UK Census – Context (2001)

- Random record swapping
- Small Cell Adjustment (SCA)
  - Lack of harmonisation
  - Some remaining risk through differencing
  - Lack of consistency between tables and a reduction in utility at low geographies
  - Late communication

<u>Winchester Local Authority:</u>

Persons in prison service establishments: 532

Number of prison service establishments: 0

UK Census – Context (2011) I

Registrars General agreement November 2006

- In line with the Statistics and Registration Service Act, 2007 (SRSA)
- More importance placed on protecting attribute disclosure than identification

- Small cells (0s, 1s, 2s) allowed provided
  - there is sufficient uncertainty that those cell counts are real,
  - and that creating this uncertainty does not cause significant damage to the data.

	Income			
Age	Low	Medium	High	Total
16-24	3	0	0	3
25-29	2	2	0	4
30-39	0	2	0	2
40-49	5	4	2	11
50-59	1	6	5	12
60-64	0	0	1	1
65+	0	0	0	0
Total	11	14	8	33

UK Census – Context (2011) II

Aims

- Harmonised approach across the UK Census Offices
- Additivity and consistency across tables
- Numbers of households, CEs and persons to remain unchanged at all geographies

- Record swapping
  - Targeted to ‘risky’ records
  - Swap records only as far as necessary

Sufficient uncertainty: Data quality

Data capture error

Elapsed time

	Primary Phobia							
Jumper Colour	Heights	Spiders	Snakes	Forms	Clowns	Other	None	Total
Black	6	8	3	0	2	2	6	27
Blue	6	7	2	0	3	1	7	26
Red	2	2	1	0	0	3	0	8
Black & White	5	9	5	0	1	3	9	32
Red & White	0	0	0	0	1	0	0	1
Black & Yellow	0	0	0	0	0	0	0	0
Other	3	7	2	0	0	4	3	19
Total	22	33	13	0	7	13	25	113

Respondent error

Sufficient uncertainty: Imputation

Non-response imputation

	Primary Phobia							
Jumper Colour	Heights	Spiders	Snakes	Forms	Clowns	Other	None	Total
Black	6	8	3	1	2	2	6	28
Blue	6	7	2	1	3	1	7	27
Red	2	2	1	0	0	4	0	9
Black & White	5	9	5	4	1	3	10	37
Red & White	0	0	0	0	1	0	0	1
Black & Yellow	0	0	0	0	0	0	0	0
Other	3	7	4	2	0	4	3	23
Total	22	33	15	8	7	14	26	125

- Record swapping
  - Targeted to ‘risky’ records
  - Swap records only as far as necessary
  - Consider imputation as part of the protection

SDC for households: Calculating the risk

Whole households are swapped

- Risk score calculated for each individual within the household.
- If an individual is rare, based on univariate distributions on a set of risk variables, the household is flagged as high risk.

Risk Variable 1: Jumper pattern
50 out of 300 persons in the OA have striped jumper
= Not risky

Risk Variable 2: Pet ownership
1 out of 300 persons in the OA own a blue dog
= Individual is risky

Household flagged as a risky household at OA level

SDC for households: Sample selection I

- Swap rate of the DG inversely related to the level of non-response
- The swap rate for DGs will lie within a set range

The swap rate across OAs is non-uniform.

- Inversely related to the OA size
- Positively related to the percentage of high risk households in the OA
- The swap rate for OAs will lie within a set range

This is part of the targeting process

SDC for households: Sample selection II

The probability of a house being sampled is affected by

- the level of non-response within the DG,
- the size of the OA and the percentage of high risk households within it, and
- whether the household has been flagged as high risk.
- All real households have a non-zero probability of being sampled

Low non-response rate in DG = Higher swap rate

Large OA = Decreases swap rate

Large number of high risk households in the OA = Increases swap rate

Household flagged as high risk = Higher chance of being sampled

SDC for households: Finding a match

The house is part of the sample

- Match only as far as necessary
- Match on household size and a set of matching variables

OA 1

Household matched on:

No. of adults

No. of children

Are there pets?

OA 2

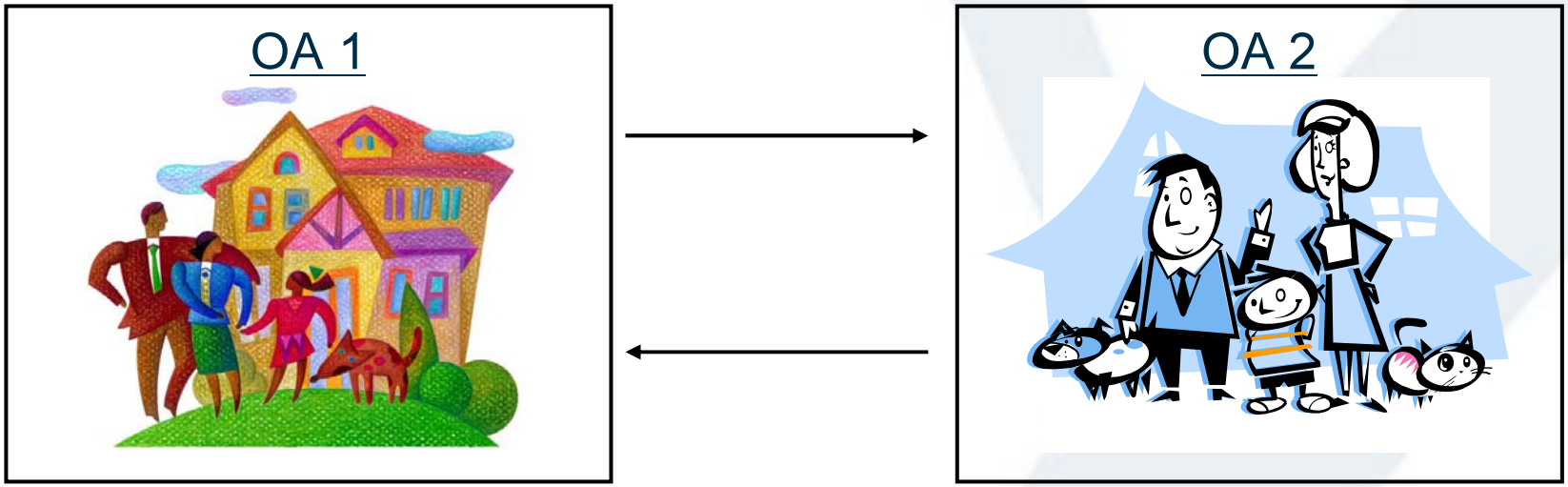
Household unique at OA, but not the MSOA

= Find match outside OA , but within the MSOA

SDC for households: Swapping

The geography variables are swapped

- What if a match isn't found?
- What about the damage to the data?



SDC for households: Recap

Households

Whole households are swapped

Risk Score

- Risk scores calculated for each individual for each risk variable
- Household is flagged as high risk if an individual is flagged as high risk

Probability

- Inversely related to the DG imputation rate
- Affected by the size of OA and distribution of high risk households
- Increased probability if flagged as high risk

Matching

- Look for matches only as far as necessary (Minimum swap: OA)
- Match on household size and other variables if possible

SDC for CEs: The rules

1. SDC methodology for CEs to remain consistent with that of households
   - Targeted record swapping

2. Keep the numbers of persons and the numbers of CEs unchanged at all geographies
   - Individual records swapped

3. Keep swapping within delivery group

SDC for CEs: The challenge

The wide range of CE types and resident types

- Population characteristics will vary between CE types and resident types
- The risk and impact of disclosure will vary between CE types and resident types

The public nature of CEs

- If a CE is identified it can essentially be viewed as a smaller geography

SDC for CEs: The aims

Maximise utility / Minimise damage

- Minimise swap rate
- Create an efficient matching process
  - Swap individuals within the same CE type
  - Swap individuals within the same resident type

    (eg. staff with staff, clients with clients, family residents with family residents)

Minimising the swap rate I

- Response rates are likely to vary as much between CE types as delivery group
- Impact and likelihood of disclosure varies between CE types

The factors which will affect the disclosure risk are:

- Rarity of CE type in the area
- Number of residents in the CE type
- Other factors impacting on uncertainty

- Set swap rate for each CE type in each MSOA

Minimising the swap rate II

- Numbers of clients and staff vary within CE types

Set swap rate for...

- staff and client residents, in each
- CE type, in each
- MSOA

- Family residents have set swap rate within the delivery group

Calculating the protection scores

For client residents:

	A	B	C	D1	E
Score	CE Type count in MSOA	Is the CE unique within the LAD?	Is the CE Type high impact?	Number of clients within the CE Type	Client Turnover
4	-	-	-	Low	-
3	Low	-	-	Med	-
2	Med	Yes	Yes	High	Low
1	High	No	No	V.High	High
0	-	-	-	0	-

Client swap rate =	0%	if CPS is	0
	X%		1-5
	Y%		6-25
	Z%		26+

For staff residents:

	A	B	C	D2
Score	CE Type count in MSOA	Is the CE unique within the LAD?	Is the CE Type high impact?	Number of staff within the CE Type
3	Low	-	-	-
2	Med	Yes	Yes	Low
1	High	No	No	High
0	-	-	-	0

Staff swap rate =	0%	if SPS is	0
	X%		1-5
	Y%		6-11
	Z%		12+

Swap within CE type

- Characteristics of residents will be different between CE types
- Swap within CE type

The problem:

- Rule 2: Keep swapping within delivery group
- How do we swap individuals in a CE type, unique in the delivery group?

- Must swap between CE types when this happens
- Matching variables chosen so key attributes remain consistent with the CE type

Swap within resident type

- Swap rates may not be the same
- Characteristics of staff, clients and family residents will be different
- Swap within resident type

- Matching variables chosen so key attributes remain consistent with the CE type
- Matching variables will be different between staff, client and family residents

Example 1: Calculating the CPS

- Resident type: Clients
- 73 client residents
- CE type: Hotel, guest house, B&B and youth hostel
- 8 CEs of this type in MSOA 1

MSOA 1

Protection score:

A = 1

B = 1

C = 1

D = 2

E = 1

1 x 1 x 1 x 2 x 1 = 2

So, CPS = 2

	A	B	C	D1	E
Score	CE Type count in MSOA	Is the CE unique within the LAD?	Is the CE Type high impact?	Number of clients within the CE Type	Client Turnover
4	-	-	-	Low	-
3	Low	-	-	Med	-
2	Med	Yes	Yes	High	Low
1	High	No	No	V.High	High
0	-	-	-	0	-

Example 1: Setting the swap rate

- Resident type: Clients
- 73 client residents
- CE type: Hotel, guest house, B&B and youth hostel
- 8 CEs of this type in MSOA 1

Protection score:

A = 1
B = 1
C = 1
D = 2
E = 1
1 x 1 x 1 x 2 x 1 = 2
So, CPS = 2

Low swap rate

MSOA 1

Client swap rate =	0%	if CPS is	0
	X%		1-5
	Y%		6-25
	Z%		26+

Example 1: Sample selection

Individuals are swapped

- Risky records are targeted
- Swap rate dependent on Protection Score

MSOA 1

High risk

= Higher chance of being sampled

Low protection score

= Lower swap rate

Example 1: Finding a match

Individual is sampled

- Matched only as far as necessary
- Matched on CE type, resident type and client specific variables

MSOA 1

Clients matched on:

Pattern of jumper

Do they have a hat?

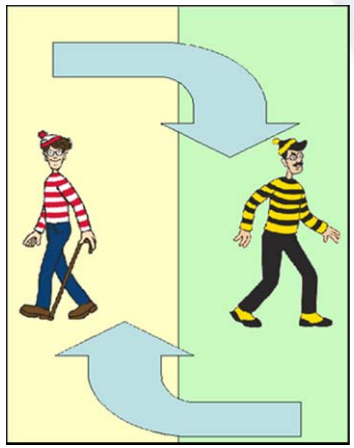
Do they have glasses?

MSOA 2

Example 1: Swapping

Individual is swapped

MSOA 1



MSOA 2

Example 2: Finding a match

Prison is unique within delivery group

= Swap individual outside of the CE type

- Matched only as far as necessary
- Matched on resident type and client specific variables

MSOA 1

Clients matched on:

Pattern of jumper

Do they have a hat?

Do they have glasses?

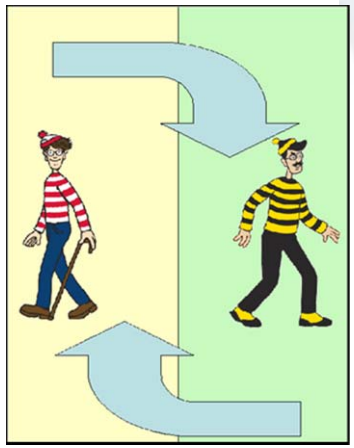
MSOA 2

Example 2: Swapping

Individual is swapped

- Still able to find a match
- Limit the damage to the data

MSOA 1



MSOA 2

SDC for households and CEs: Recap

Households

Whole households are swapped

Risk Score

- Risk scores calculated for each individual for each risk variable
- Household is flagged as high risk if an individual is flagged as high risk

Probability

- Inversely related to the DG imputation rate
- Affected by the size of OA and distribution of high risk households
- Increased probability if flagged as high risk

Matching

- Look for matches only as far as necessary (Minimum swap: OA)
- Match on household size and other variables if possible

Communal Establishments

Individuals are swapped

Risk Score

- Risk score calculated for each individual for each risk variable
- Protection score calculated from CE type and resident type within the MSOA

Probability

- Positively related to protection score
- Positively related to individual record risk score

Matching

- Look for matches only as far as necessary (Minimum swap: MSOA)
- Matching variables dependent on resident type

Sufficient uncertainty: Record swapping

Record Swapping

	Primary Phobia							
Jumper Colour	Heights	Spiders	Snakes	Forms	Clowns	Other	None	Total
Black	6	8	3	1	2	2	6	28
Blue	6	7	2	1	3	1	7	27
Red	2	2	1	0	0	4	0	9
Black & White	5	9	5	4	1	3	10	37
Red & White	0	0	0	0	0	0	0	0
Black & Yellow	0	0	1	0	0	0	0	1
Other	3	7	4	2	0	4	3	23
Total	22	33	16	8	6	14	26	125

Sparsity rule

Reducing the detail reduces the risk of disclosure

	Primary Phobia							
Jumper Colour	Heights	Spiders	Snakes	Forms	Clowns	Other	None	Total
Plain	14	17	6	2	5	7	13	64
Striped	5	9	6	4	1	3	10	38
Other	3	7	4	2	0	4	3	23
Total	22	33	16	8	6	14	26	125
Red & White	0	0	0	0	0	0	0	0
Black & Yellow	0	0	1	0	0	0	0	1
Other	3	7	4	2	0	4	3	23
Total	22	33	16	8	6	14	26	125

Sparsity rule

- Internal cells must have an average cell value > n

(Table population total ÷ Number of internal cells) > n

Threshold rule

A smaller population has a greater risk of disclosure

	Primary Phobia							
Jumper Colour	Heights	Spiders	Snakes	Forms	Clowns	Other	None	Total
Black	6	8	3	1	2	2	6	28
Blue	6	7	2	1	3	1	7	27
Red	2	2	1	0	0	4	0	9
Black & White	5	9	5	4	1	3	10	37
Red & White	0	0	0	0	0	0	0	0
Black & Yellow	0	0	1	0	0	0	0	1
Other	3	7	4	2	0	4	3	23
Total	22	33	16	8	6	14	26	125

- OAs have a minimum size threshold of 100 persons and 40 households

Where else can we use thresholds?

- Minority/Small populations

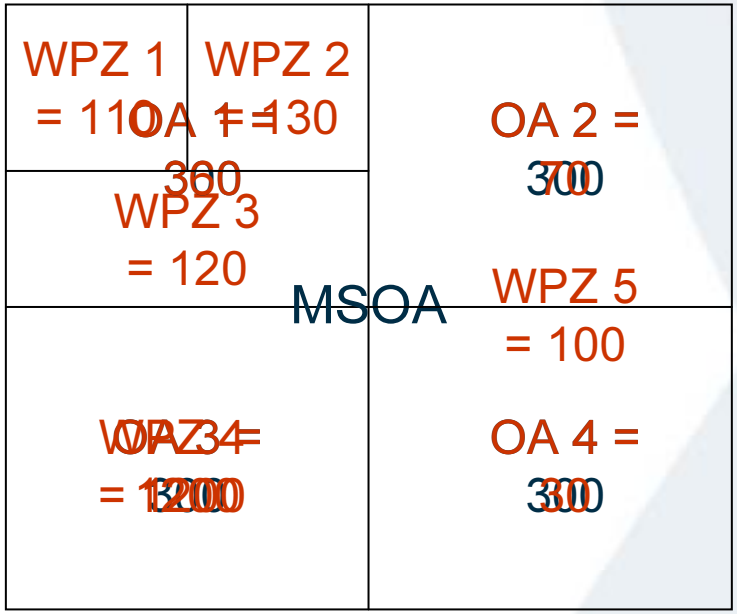
Workplace zones

- Produce statistics based on the workplace population
- The workplace populations will vary between OA

- Merge and split OAs to create ‘workplace zones’
- Keep workplace zones nested within MSOA

Population Sizes

Workplace Sizes



Calculating uncertainty I

Original table

Non-response imputation

	Primary Phobia							
Jumper Colour	Heights	Spiders	Snakes	Forms	Clowns	Other	None	Total
Black	6	8	3	1	2	2	6	28
Blue	6	7	2	1	3	1	7	27
Red	2	2	1	0	0	4	0	9
Black & White	5	9	5	4	1	3	10	37
Red & White	0	0	0	0	0	0	0	0
Black & Yellow	0	0	1	0	0	0	0	1
Other	3	7	4	2	0	4	3	23
Total	22	33	16	8	6	14	26	125

- Require a measure of uncertainty

Calculating uncertainty II

	Primary Phobia							
Jumper Colour	Heights	Spiders	Snakes	Forms	Clowns	Other	None	Total
Black	6	8	3	1	2	2	6	28
Blue	6	7	2	1	3	1	7	27
Red	2	2	1	0	0	4	0	9
Black & White	5	9	5	4	1	3	10	37
Red & White	0	0	0	0	0	0	0	0
Black & Yellow	0	0	1	0	0	0	0	1
Other	3	7	4	2	0	4	3	23
Total	22	33	16	8	6	14	26	125

Success Measure 1

Proportion of real cases of attribute disclosure that have been protected, either by

- Record swapping, or
- Non-response imputation

Calculating uncertainty III

	Primary Phobia							
Jumper Colour	Heights	Spiders	Snakes	Forms	Clowns	Other	None	Total
Black	6	8	3	1	2	2	6	28
Blue	6	7	2	1	3	1	7	27
Red	2	2	1	0	0	4	0	9
Black & White	5	9	5	4	1	3	10	37
Red & White	0	0	0	0	0	0	0	0
Black & Yellow	0	0	1	0	0	0	0	1
Other	3	7	4	2	0	4	3	23
Total	22	33	16	8	6	14	26	125

Success Measure 2

Proportion of apparent cases of attribute disclosure that are not real, having been created either by

- Record swapping, or
- Non-response imputation

Calculating uncertainty IV

Doubt Measure

- Doubt = [1 – (1-S1)(1-S2)]

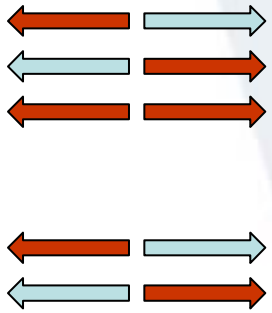
where

- S1 = Proportion of real Attribute Disclosure (AD) cases protected
- S2 = Proportion of apparent AD cases that are not real

Example: Calculating uncertainty I

Unprotected Table

	Income			
Age	Low	Medium	High	Total
16-24	3	0	0	3
25-29	2	2	0	4
30-39	0	2	0	2
40-49	5	4	2	11
50-59	1	6	5	12
60-64	0	0	1	1
65+	0	0	0	0
Total	11	14	8	33



Protected Table

	Income			
Age	Low	Medium	High	Total
16-24	3	2	1	6
25-29	2	0	0	2
30-39	0	2	0	2
40-49	5	4	4	13
50-59	1	6	5	14
60-64	0	0	0	0
65+	0	1	0	1
Total	11	15	10	38

S1 = Proportion of real AD cases protected

= AD cases removed ÷ AD cases in unprotected table

= (3 + 1) ÷ (3 + 2 + 1) = 0.667

S2 = Proportion of apparent AD cases that are not real

= AD cases created ÷ AD cases in protected table

= (2 + 1) ÷ (2 + 2 + 1) = 0.6

Example: Calculating uncertainty II

Unprotected Table

	Income			
Age	Low	Medium	High	Total
16-24	3	0	0	3
25-29	2	2	0	4
30-39	0	2	0	2
40-49	5	4	2	11
50-59	1	6	5	12
60-64	0	0	1	1
65+	0	0	0	0
Total	11	14	8	33

Protected Table

	Income			
Age	Low	Medium	High	Total
16-24	3	2	1	6
25-29	2	0	0	2
30-39	0	2	0	2
40-49	5	4	4	13
50-59	1	6	5	14
60-64	0	0	0	0
65+	0	1	0	1
Total	11	15	10	38

Doubt = [1 – (1-S1)(1-S2)]

S1 = 0.667

S2 = 0.6

Doubt = [1 – (1-0.667)(1- 0.6)] = 0.867

Doubt = 86.7%

Public perception

Unprotected Table

	Income			
Age	Low	Medium	High	Total
16-24	3	0	0	3
25-29	2	2	0	4
30-39	0	2	0	2
40-49	5	4	2	11
50-59	1	6	5	12
60-64	0	0	1	1
65+	0	0	0	0
Total	11	14	8	33

Protected Table

	Income			
Age	Low	Medium	High	Total
16-24	3	2	1	6
25-29	2	0	0	2
30-39	0	2	0	2
40-49	5	4	4	13
50-59	1	6	5	14
60-64	0	0	0	0
65+	0	1	0	1
Total	11	15	10	38

- Where am I?
- Where is the protection?

User communication is crucial

- All respondents are included in outputs (somewhere)
- Greater utility than in 2001
- Record swapping has been applied

Summary

- The primary SDC methodology will use targeted record swapping
- Numbers of households, CEs and persons will remain unchanged at all geographies
- SDC methodology aims to maximise utility:
  - Minimise amount of swapping using non-response rates (for households) and protection scores (for CEs)
  - Swap only as far as necessary
  - Use an efficient matching process
- Use of sparsity and threshold rules help produce safe outputs for small populations and workplace zones
- Development of a doubt measure to assess the levels of uncertainty of an output table

Q&A

sdc.queries@ons.gsi.gov.uk