The potential role of small area estimation

Dr Paul Williamson

Dept. of Geography & Planning

UNIVERSITY OF LIVERPOOL

(1) What is Small Area Estimation?

Box F: Statistics possible using survey data

Area type	Average number of residents	1 year's data (800 threshold)	3 years' data (230 threshold)	5 years' data (130 threshold)
LA	160,000	Detailed cross-tabulations (c 200 cells)	Detailed cross-tabulations (c 500 cells)	Very detailed cross-tabulations (c 1000 cells)
MSOA	7,800	Some single variable statistics (c 10 cells)	Very simple cross-tabulations (c 30 cells)	Simple cross-tabulations (c 50 cells)
LSOA	1,600	Not available	Some single variable statistics (c 5 cells)	Some single variable statistics (c 10 cells)
OA	300	Not available	Not available	Not available

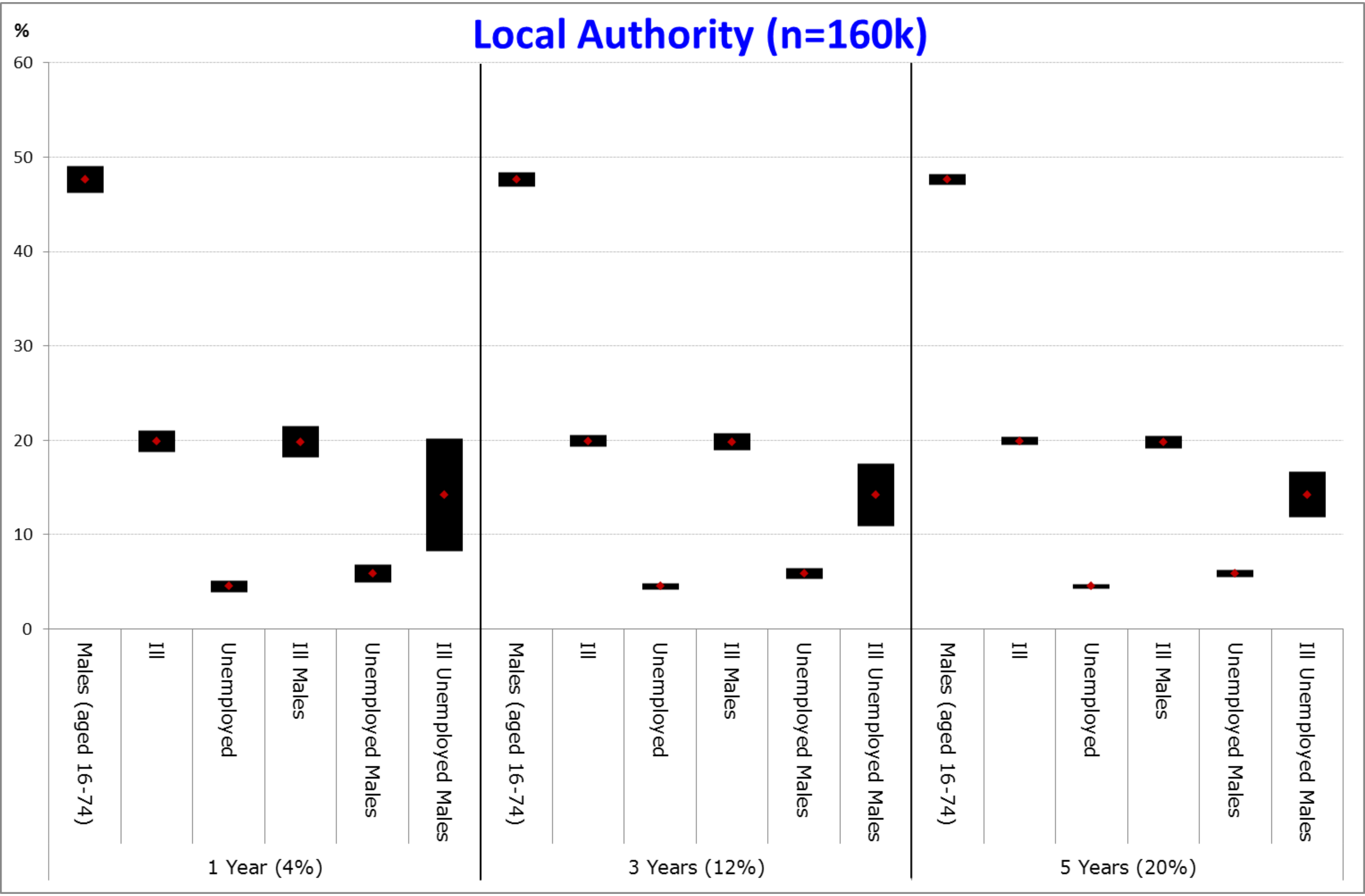
(2) Direct survey estimation

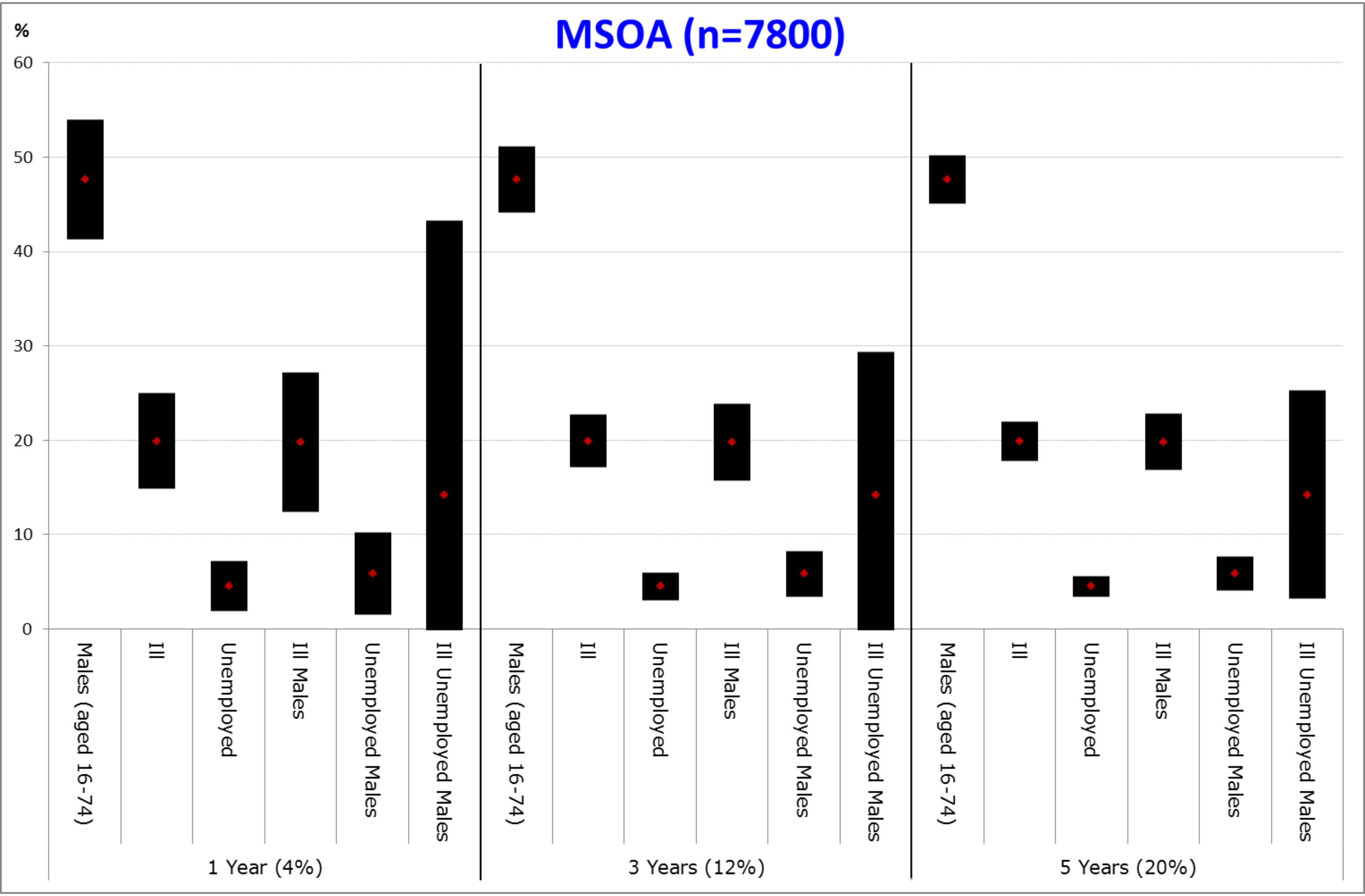
Barking & Dagenham

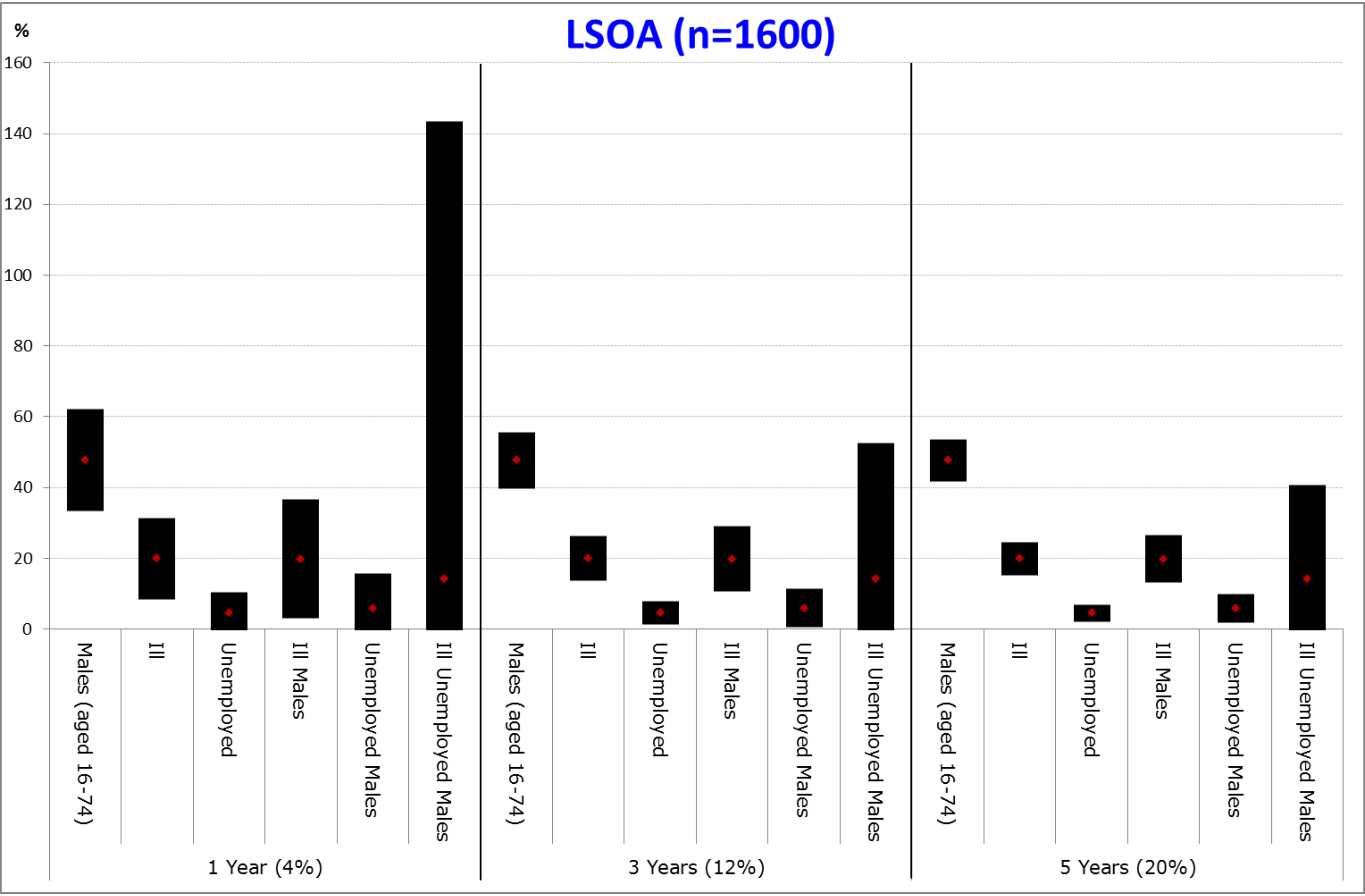
	LA
Pop. Attribute	*n*
Person	160000
Persons aged 16-74	113577
Males (aged 16-74)	54099
Ill	22638
Unemployed	5121
Ill Males	10729
Unemployed Males	3174
Ill Unemployed Males	452

Cell count < 800

Local Authority (n=160k)




MSOA (n=7800)
%
60
50
40
30
20
10
0
Males (aged 16-74)
Ill
Unemployed
Ill Males
Unemployed Males
Ill Unemployed Males
Males (aged 16-74)
Ill
Unemployed
Ill Males
Unemployed Males
Ill Unemployed Males
Males (aged 16-74)
Ill
Unemployed
Ill Males
Unemployed Males
Ill Unemployed Males
1 Year (4%)
3 Years (12%)
5 Years (20%)



LSOA (n=1600)
%
160
140
120
100
80
60
40
20
0
Males (aged 16-74)
Ill
Unemployed
Ill Males
Unemployed Males
Ill Unemployed Males
1 Year (4%)
3 Years (12%)
5 Years (20%)

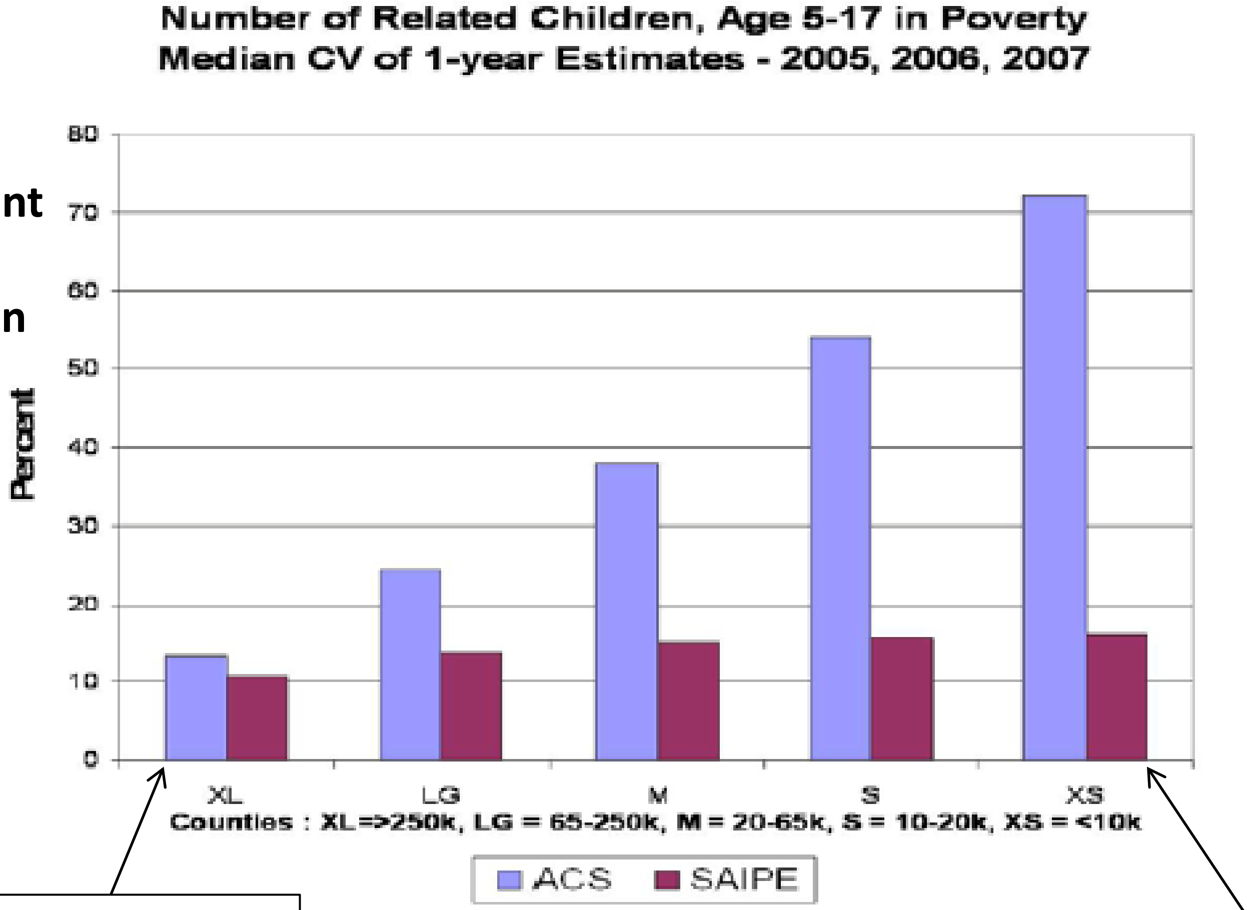
(3) Other main SAE approaches

Proxies

Ecological regression

- Find relationship between AREA-level **Y** and **X**(s) for areas sampled in survey
- Assume applies to (non-sampled) areas, for which AREA-level X is known
- E.g. ONS small area income estimates for MSOAs

Coefficient of Variation



Big areas
(pop > 250k)

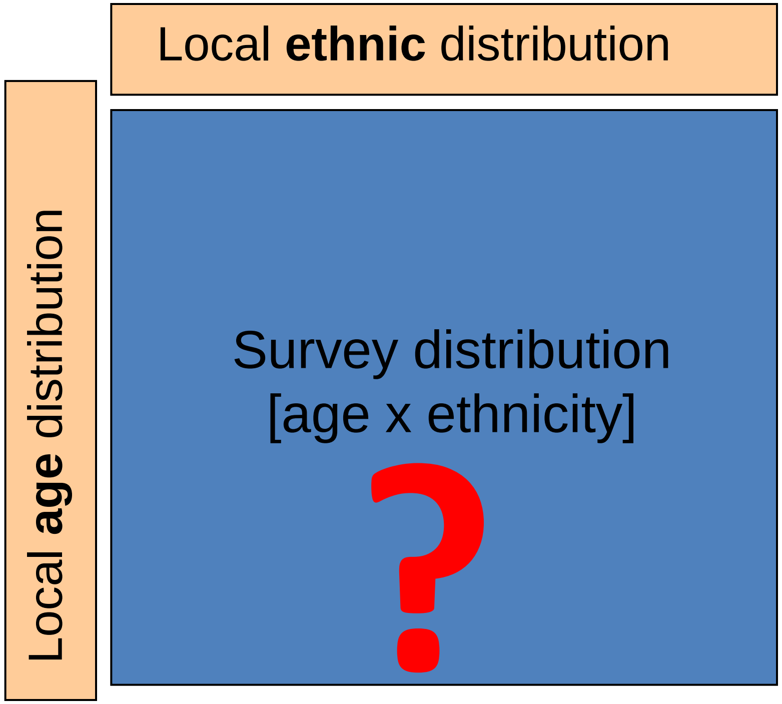
Small areas
(median pop. c 1k)

Known problems with Ecological Regression

- Regression to the mean
- Point estimate
- Covariate dependent

Survey reweighting / calibration

Reweight survey data to fit local area constraints/margins...



Local **ethnic** distribution

Local **age** distribution

Survey distribution
[age x ethnicity]

?

...potentially weighting DOWN instead of up

Reweighting approaches:

- IPF / raking/ Mostellerisation / Cross-Fratar / RAS etc../
- Generalised Regression (GREG)
- Integer linear programming solved using simplex or integer point methods
- etc…

Known problems with reweighting approaches:

- As per ecological regression…
- **BUT** provides distributional rather than point estimates

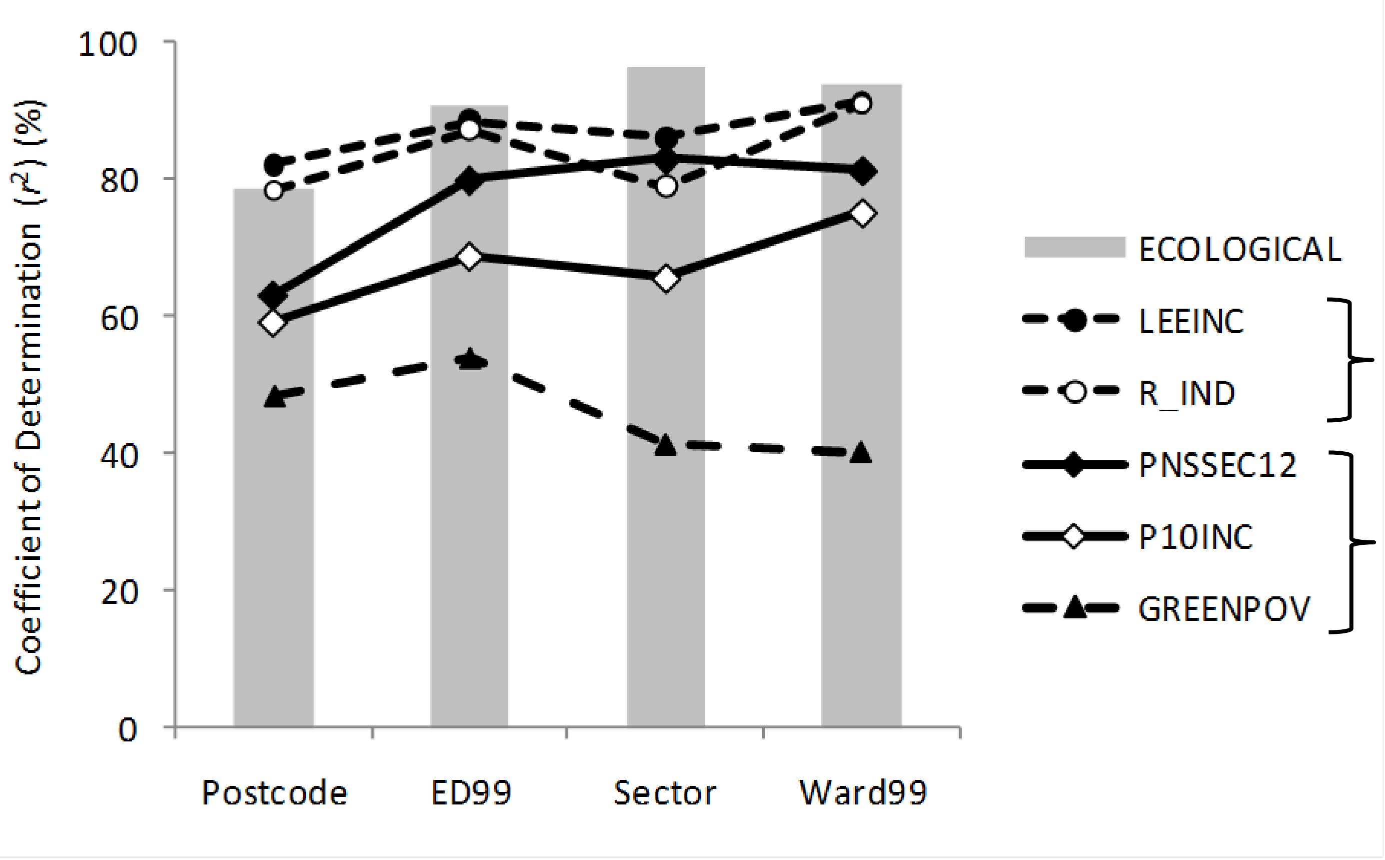
Record-level imputation

- Impute (estimate) missing data onto existing record level data
- E.g. Impute income onto Census records given known individual attributes such as age and occupation

Known problems:

- As per ecological regression, plus:
- Requires record-level data with 100% local area coverage
- **BUT** does provide distributional rather than point estimates

(4) Case study: mean income

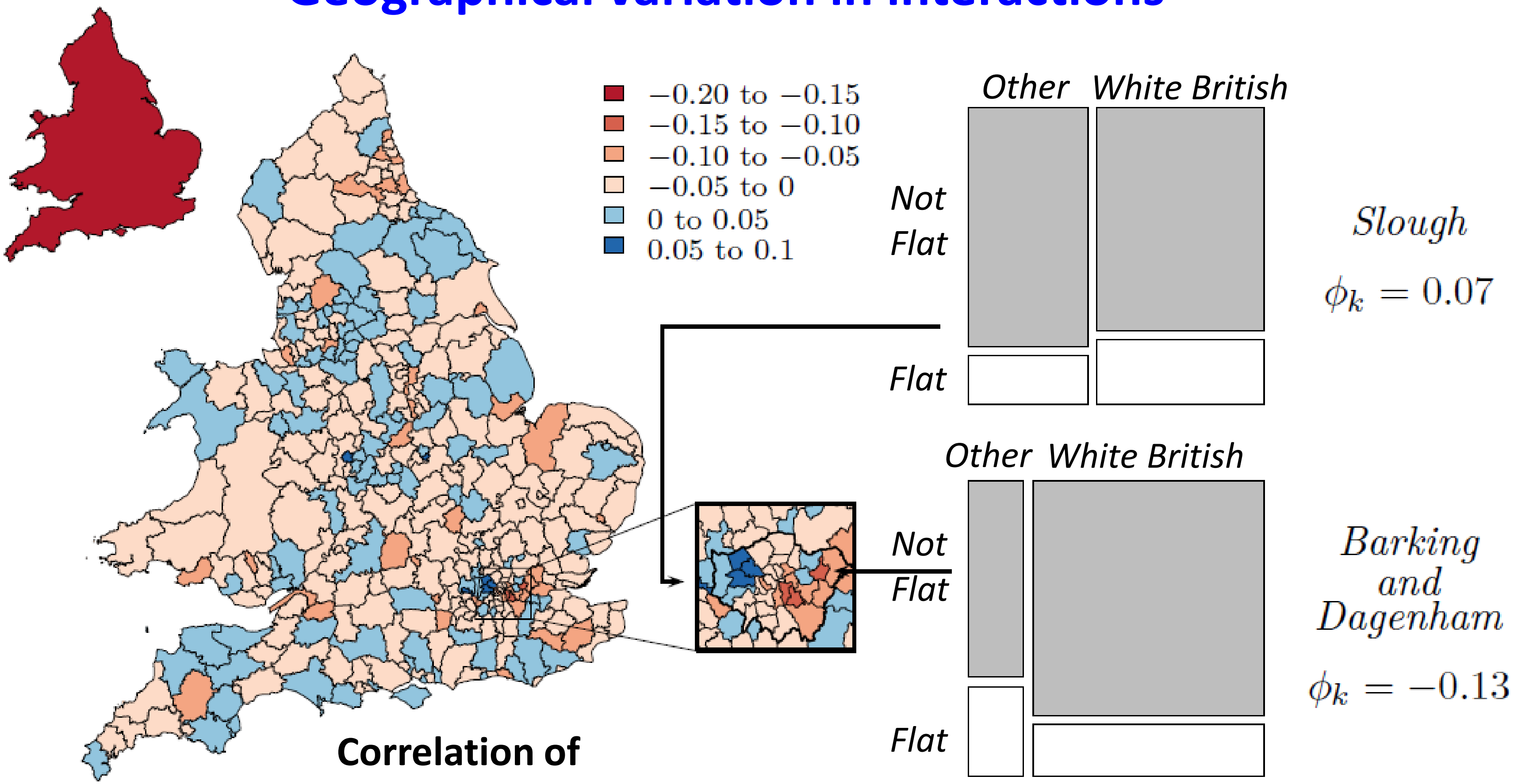


Imputation

Proxies

(5) The limitations of SAE

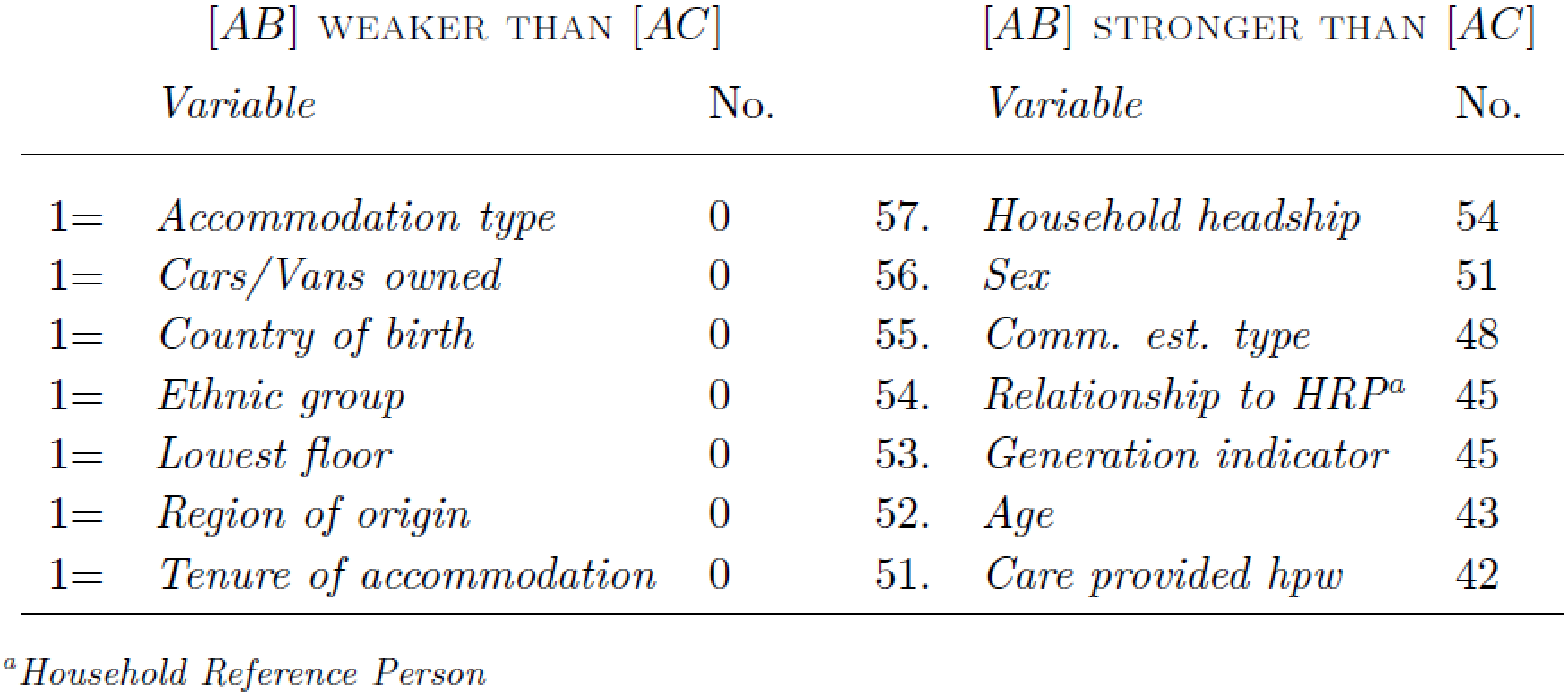
Geographical variation in interactions



Correlation of *Accommodation type* with *Ethnicity*

Geography MORE important (Top 7)

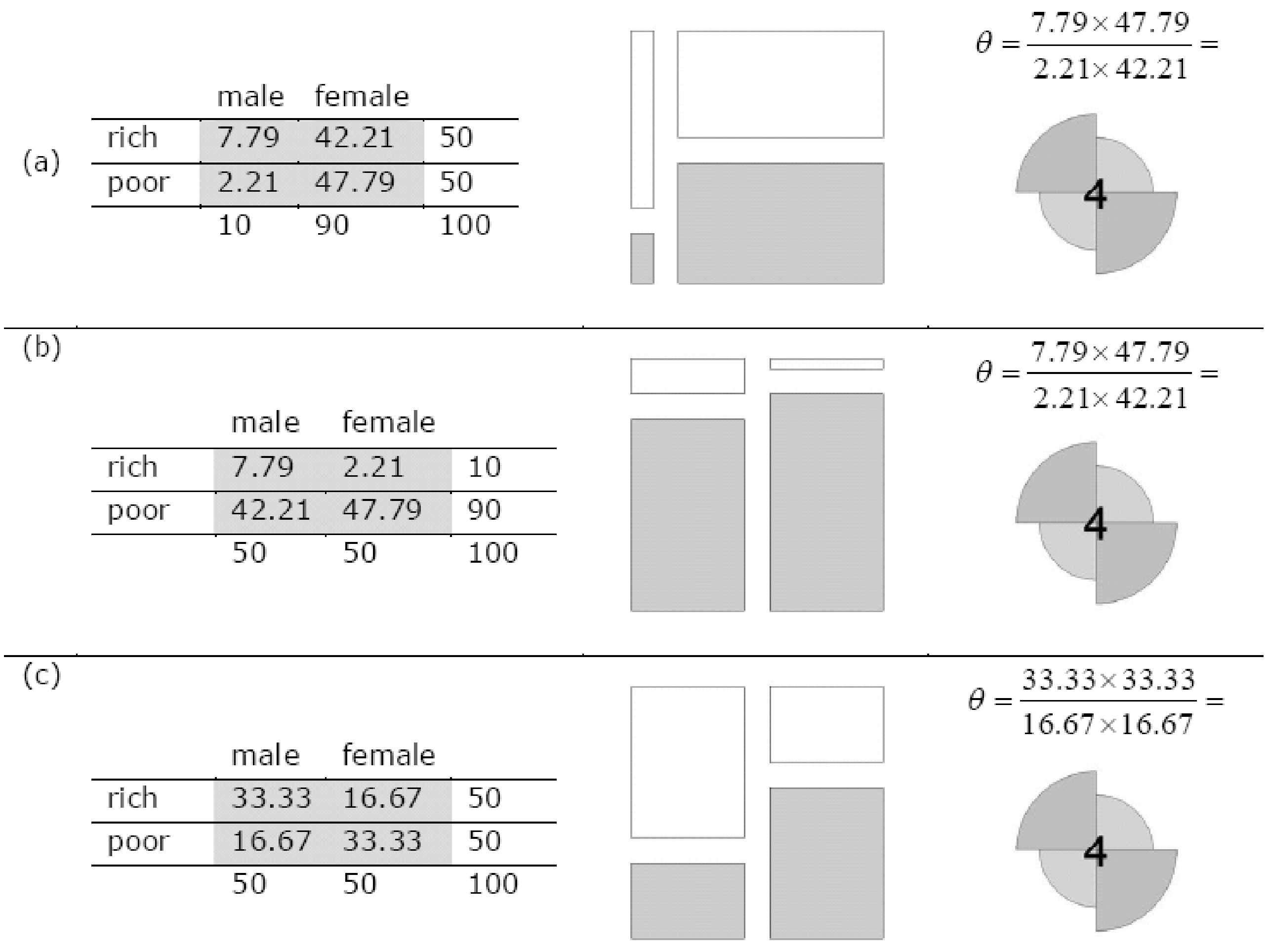
Geography LESS important (Bottom 7)



	[*AB*] WEAKER THAN [*AC*]			[*AB*] STRONGER THAN [*AC*]	
	Variable	No.		*Variable*	No.
1=	*Accommodation type*	0	57.	*Household headship*	54
1=	*Cars/Vans owned*	0	56.	*Sex*	51
1=	*Country of birth*	0	55.	*Comm. est. type*	48
1=	*Ethnic group*	0	54.	*Relationship to HRP*[a]	45
1=	*Lowest floor*	0	53.	*Generation indicator*	45
1=	*Region of origin*	0	52.	*Age*	43
1=	*Tenure of accommodation*	0	51.	*Care provided hpw*	42

[a] *Household Reference Person*

For the perfect estimate, need to know margins AND interactions



(a)

	male	female	
rich	7.79	42.21	50
poor	2.21	47.79	50
	10	90	100

$$\theta = \frac{7.79 \times 47.79}{2.21 \times 42.21} =$$

(b)

	male	female	
rich	7.79	2.21	10
poor	42.21	47.79	90
	50	50	100

$$\theta = \frac{7.79 \times 47.79}{2.21 \times 42.21} =$$

(c)

	male	female	
rich	33.33	16.67	50
poor	16.67	33.33	50
	50	50	100

$$\theta = \frac{33.33 \times 33.33}{16.67 \times 16.67} =$$

FIGURE 4- *Tabular and graphical displays of three possible marginal distributions (a), (b) and (c) with the same odds ratio = 4.*

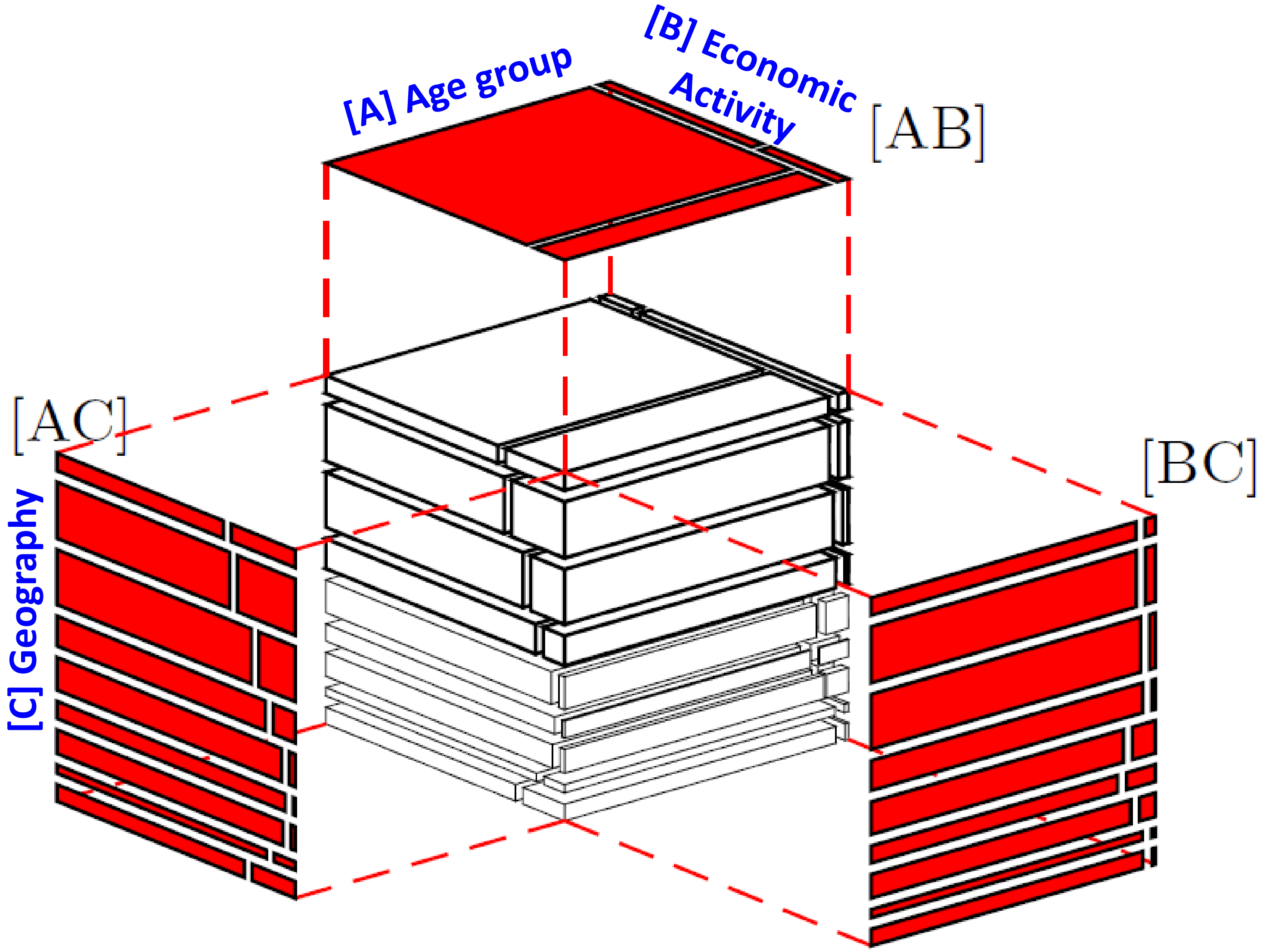
(6) SAE implications of Admin data + Survey approach

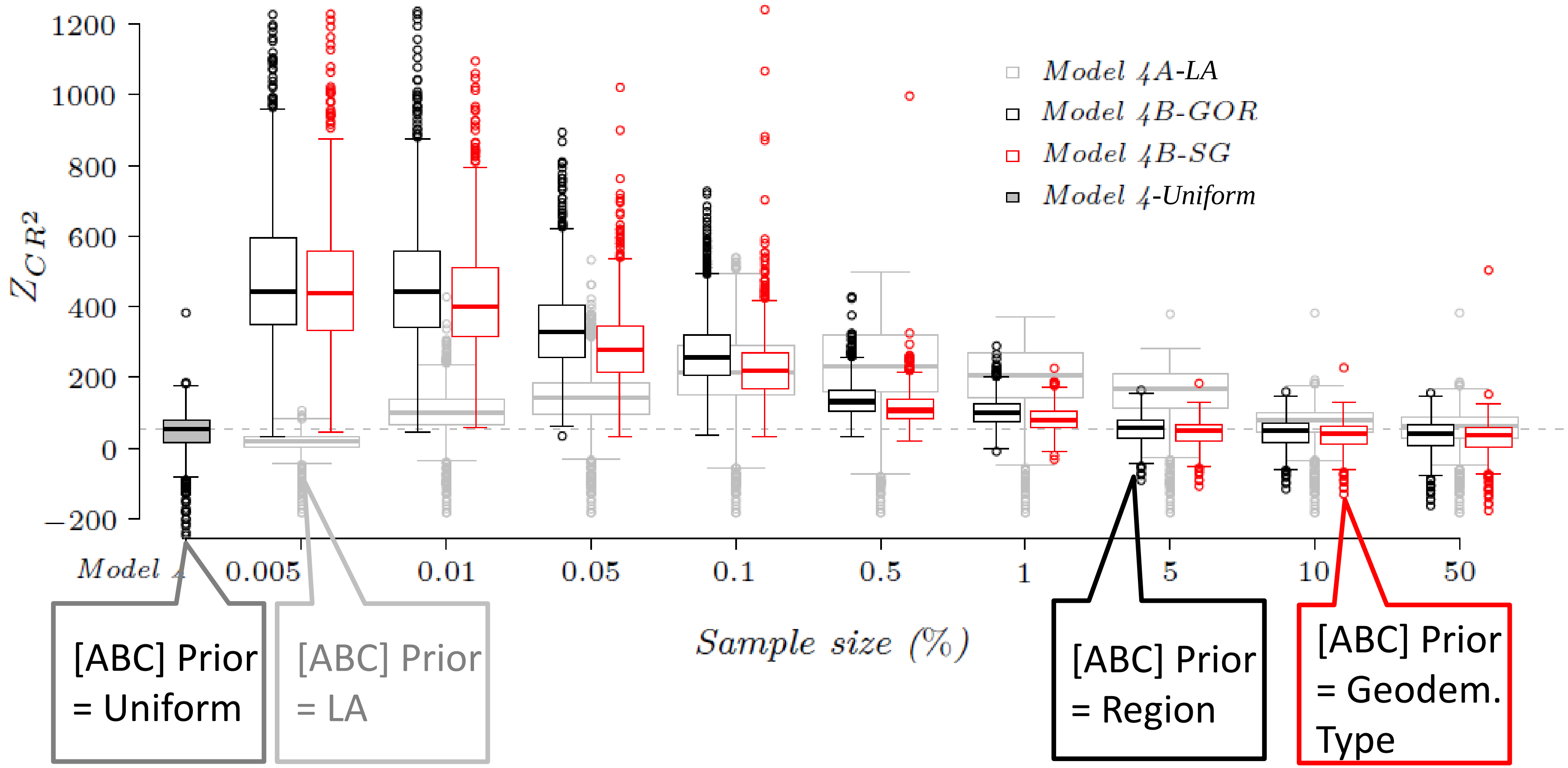
- Sample Survey data for ALL areas, not just some

BUT

- Sampled local area interactions unreliable
- No census ➔ few reliable covariates
- No census ➔ validation of covariates?
- ONS/user SAE workload
- Unavoidable regression to the mean
- SAE reliability that varies by topic and geographic scale

Supplementary slides





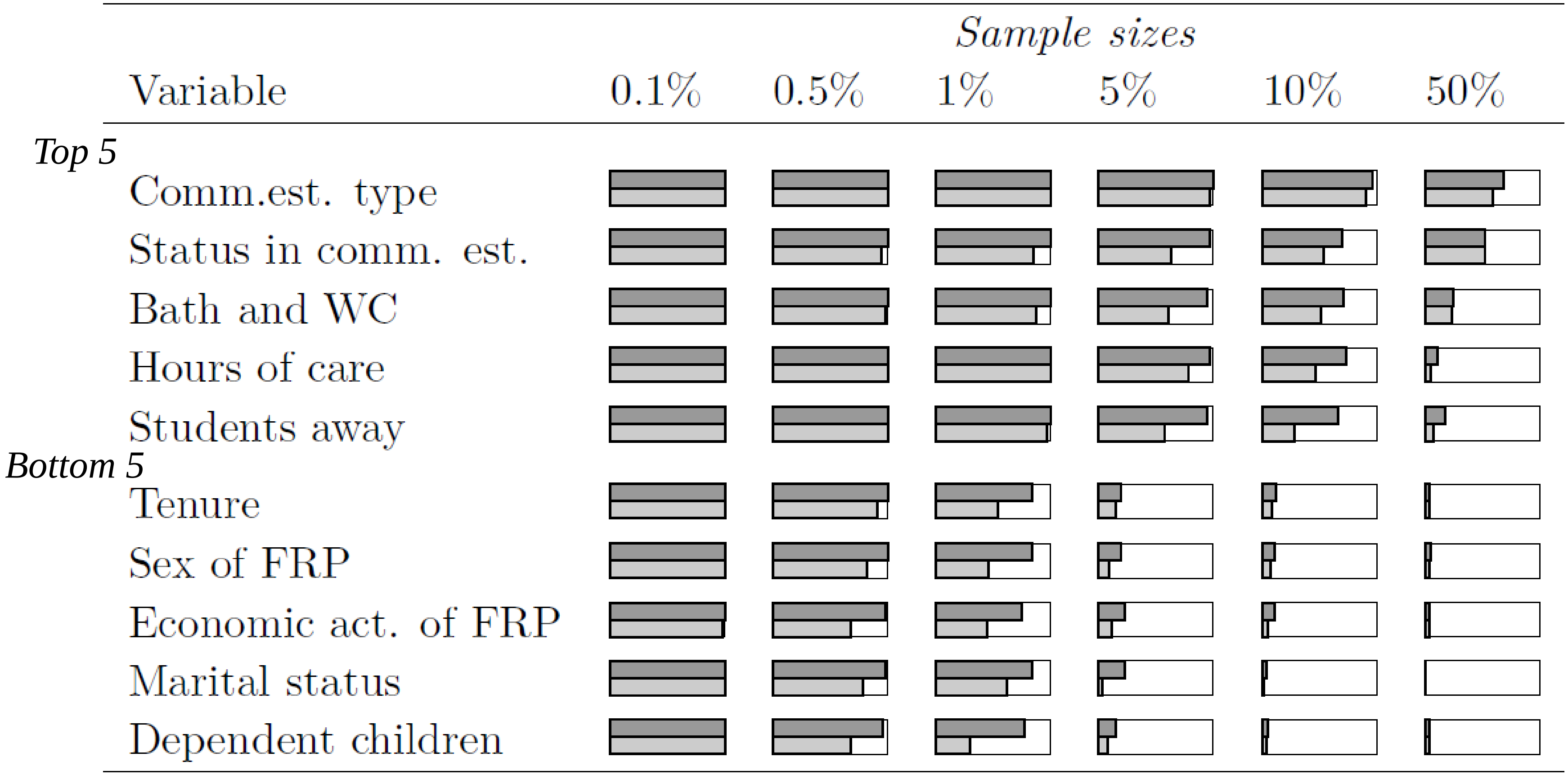
[ABC] Prior = Uniform

[ABC] Prior = LA

[ABC] Prior = Region

[ABC] Prior = Geodem. Type

Table 11.3: Proportion of tables where *Uniform [ABC]* outperforms regional (darker grey) and Supergroup sampling (lighter grey) - top and bottom five variables



Cross-level regression

- Find relationship between ***INDIVIDUAL-level*** **Y** and **X**(s) in survey
- Assume applies to non-sampled areas, for which ***AREA-level*** **X** is known
- E.g. Estimates of local area ‘Breadline poverty’ and ‘Fuel poverty’ rates

Known problems:

- As per ecological regression plus..
- Commits Ecological fallacy
- Ecological regression performs better (when possible)

		Surrogate/Estimate			
		% NSSEC 1+2	Individual Regression	Sub-group mean	Ecological Regression
		[PNSSEC12]	[R_IND]	[LEEINC]	[ECOLOGICAL]
		% ranked in same decile as income			
Decile	1	71	66	74	80
[low income]	2	46	34	40	52
	3	32	40	35	43
	4	32	26	37	40
	5	25	34	39	37
	6	17	28	45	30
	7	26	28	43	31
	8	23	35	48	46
	9	28	51	57	60
[high income]	10	55	77	82	82
	Overall	*36*	*42*	*50*	46
	Within ± 1 decile	*82*	*84*	*89*	*92*