



# Imputing Socioeconomic Status onto Administrative Datasets

2nd General Conference of the  
International Microsimulation Association, Ottawa

L Brown, L Thurecht, A Armstrong, C Gong, A Harding  
NATSEM  
10 June 2009

# Background

- Administrative data is rich with details related to the particular service being provided or activity being recorded
- However, such records typically contain little, if any, details relating to the socioeconomic status (SES) of the people to which they relate
- Yet there is typically a lot of policy interest in the distributional characteristics of users of a service:
  - Understand who is (not) benefiting from a service
  - Better targeting of a service
  - Scope for alternative policy settings

# Areal Measures of Socioeconomic Status

- A common approach to addressing this problem is to assign an areal-based measure of the variable of interest:
    - ABS SEIFAs (socioeconomic indexes for areas) are perhaps the most commonly used indicator in Australia
    - In particular, they are typically applied at the SLA level
    - In June 2005:
      - 97 SLAs had a population greater than 50,000
      - the SLA with the largest population was 192,000
- ⇒ *One measure to represent the characteristics of every person within that area*

# Ecological Fallacy

- Knowing something in aggregate about an area does not enable you to infer anything about an individual within that area

# Inferring Person Level SES from Areal Data

## - The Basic Idea

- Replicate the known distribution of a variable of interest on a set of records that does not contain the variable of interest (in this case, SES):
  - Identify the donor distribution eg using the ABS Census of Population and Housing (the empirical distribution)
  - Select matching variables between the source of the empirical distribution and the administrative records
  - For each unique set of matched variables, randomly assign a value of the variable of interest on the administrative records in such a way as to replicate the empirical distribution

# A Hypothetical Example

- Say in a particular areal unit there are twenty 0-10 year old males with half being in the lowest SES group and half in the highest SES group
- Say we also have a set of administrative records that show four 0-10 year old males from the same areal unit
- Impute SES onto the administrative records by randomly assigning half to the lowest SES group and half to the highest SES group

# A Hypothetical Example (cont)

## Empirical Distribution (Census):

Area	Sex	Age	SES Quintile	Number of People	Proportion of Area	Cumulative Proportion
123456	Male	0-10	1	10	0.5	0.5
123456	Male	0-10	2	0	0.0	0.5
123456	Male	0-10	3	0	0.0	0.5
123456	Male	0-10	4	0	0.0	0.5
123456	Male	0-10	5	10	0.5	1.0
				$\Sigma = 20$		

## Administrative Data:

Area	Sex	Age	Person ID	$z^*$	Imputed SES Quintile
123456	Male	0-10	438151	0.4140112	1
123456	Male	0-10	68656	0.6239384	5
123456	Male	0-10	998134	0.7950528	5
123456	Male	0-10	3767852	0.3979469	1

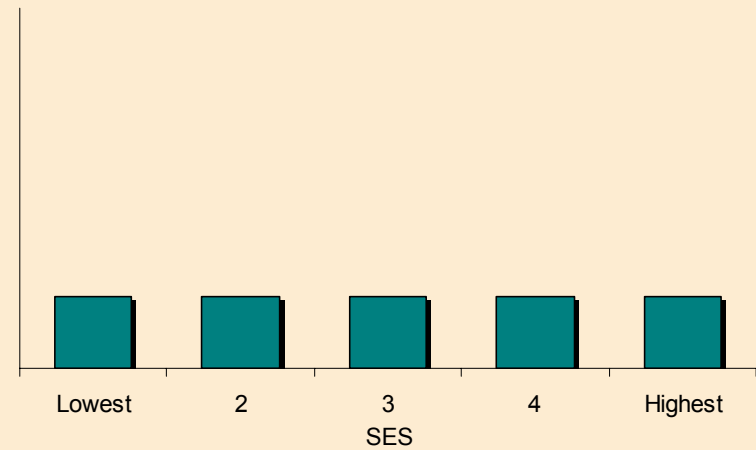
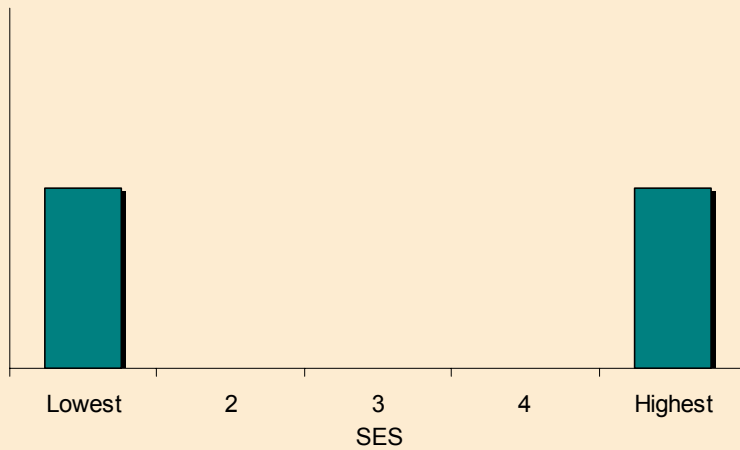
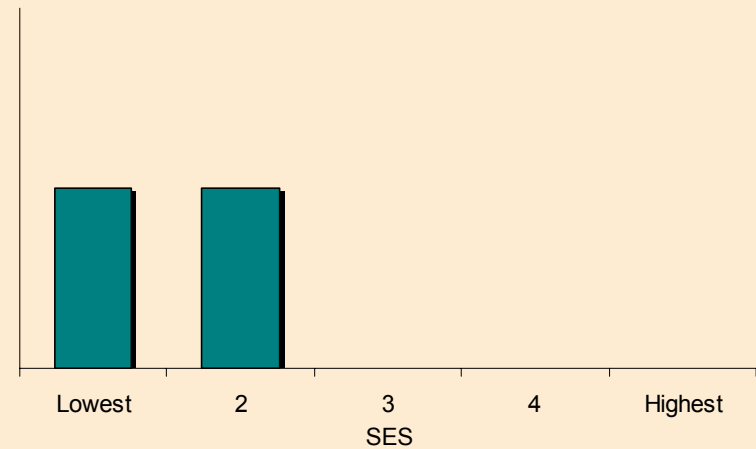
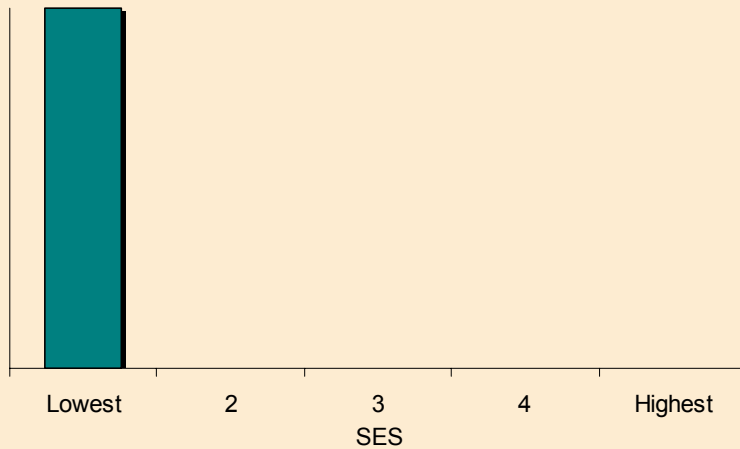
\* where  $z \sim U(0,1)$

# Fine Grained Distributions vs Confidentiality

- Inverse relationship between the size of areal unit and homogeneity of the population
  - Ideally want to obtain the most detailed distribution possible (ie high number of unique combinations among matching variables):  
⇒ Implies more homogenous sub-populations
- However, greater likelihood that an individual could then be identified from the empirical distribution
- Trade-off between more detailed homogenous sub-populations and a potential loss of confidentiality associated with the donor empirical distribution:
  - Australian Bureau of Statistics process of randomly perturbing cells

# Assumption of Independence and Potential Bias

## Potential Empirical Distributions



# Potential Sources of Error

- Random imputation error:
  - Different seed for the random number generator → different random number → different imputation
  - Empirical distribution is retained (as  $n$  becomes large), but the individually assigned SES may be different
- Pattern of use in the administrative data is not identical to the population based empirical distribution:
  - Modify the imputation based on:
    - Exogenous information?
    - Informed assumption?

# Quantitative Assessment of Imputed Distribution

## Case 1: Heterogeneous Population

	P( $\hat{S}ES = 1$ ) = 0.2	P( $\hat{S}ES = 2$ ) = 0.2	P( $\hat{S}ES = 3$ ) = 0.2	P( $\hat{S}ES = 4$ ) = 0.2	P( $\hat{S}ES = 5$ ) = 0.2
P( $SES = 1$ ) = 0.2	0.04	0.04	0.04	0.04	0.04
P( $SES = 2$ ) = 0.2	0.04	0.04	0.04	0.04	0.04
P( $SES = 3$ ) = 0.2	0.04	0.04	0.04	0.04	0.04
P( $SES = 4$ ) = 0.2	0.04	0.04	0.04	0.04	0.04
P( $SES = 5$ ) = 0.2	0.04	0.04	0.04	0.04	0.04

$$P(SE S = \hat{S}ES) = 0.04 * 5 = 0.2$$

$$P(SE S <> \hat{S}ES) = 0.04 * 20 = 0.8$$

## Case 2: Homogeneous Population

	P( $\hat{S}ES = 1$ ) = 0.2	P( $\hat{S}ES = 2$ ) = 0.2	P( $\hat{S}ES = 3$ ) = 0.2	P( $\hat{S}ES = 4$ ) = 0.2	P( $\hat{S}ES = 5$ ) = 0.2
P( $SES = 1$ ) = 0.9	0.81	0	0	0	0.09
P( $SES = 2$ ) = 0.0	0	0	0	0	0
P( $SES = 3$ ) = 0.0	0	0	0	0	0
P( $SES = 4$ ) = 0.0	0	0	0	0	0
P( $SES = 5$ ) = 0.1	0.09	0	0	0	0.01

$$P(SE S = \hat{S}ES) = 0.81 + 0.01 = 0.82$$

$$P(SE S <> \hat{S}ES) = 0.09 + 0.09 = 0.18$$

# An Applied Example

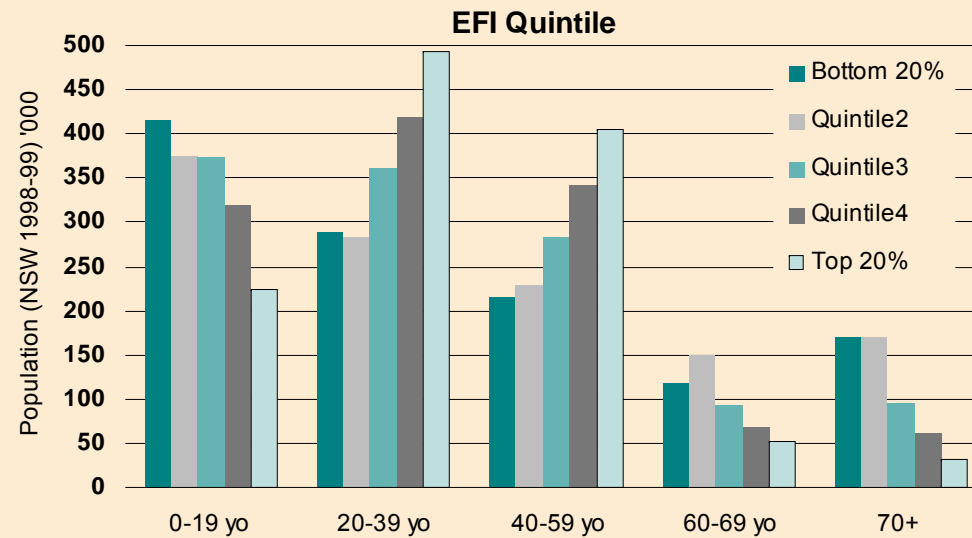
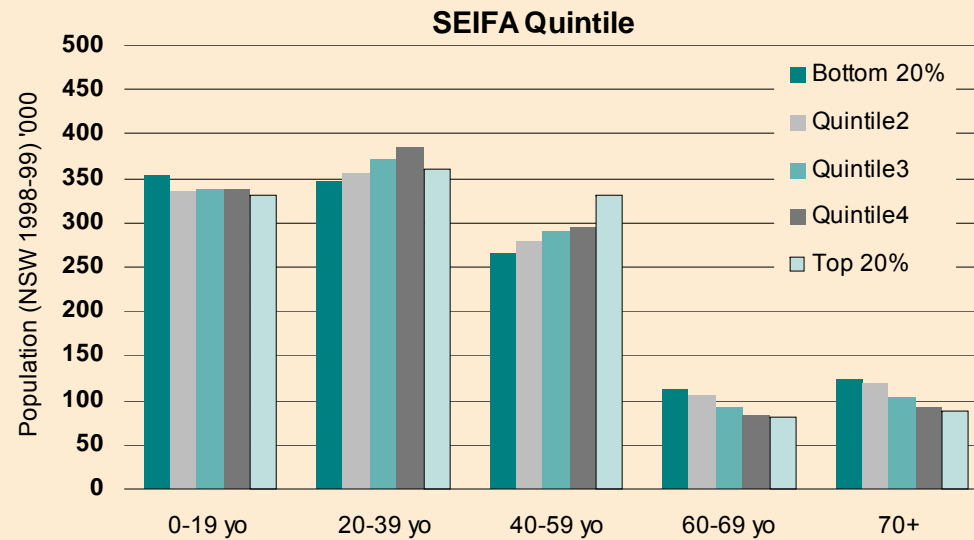
- Variable of interest is equivalent family income (EFI), taken from the Census by:
  - Census Collection district (average c 220 households in each)
  - Sex
  - Age groups (7 ten-year age groups plus one group for 70+ years)
  - EFI quintile
- The number of potential unique empirical distributions of EFI by quintile are (based on the 1996 Census for New South Wales residents):
  - $11,566 \text{ CDs} * 2 \text{ sexes} * 8 \text{ age groups} = 185,056$  distinct possible empirical distributions

# How Sparsely Populated Were the Cells?

- Around 10% of possible combinations were not represented
- Homogeneity of sub-populations:

Number of quintiles (for each CD-sex-age permutation)	CDs with $\geq 90\%$ of population located in:	CDs with $\geq 60\%$ of population located in:
1 Quintile	3.1%	11.2%
2 Quintiles	8.0%	55.8%
3 Quintiles	21.0%	33.0%
4 Quintiles	46.3%	0.0%
5 Quintiles	21.6%	0.0%

# NSW Population by Socioeconomic Status: Imputed SES vs ABS SEIFA

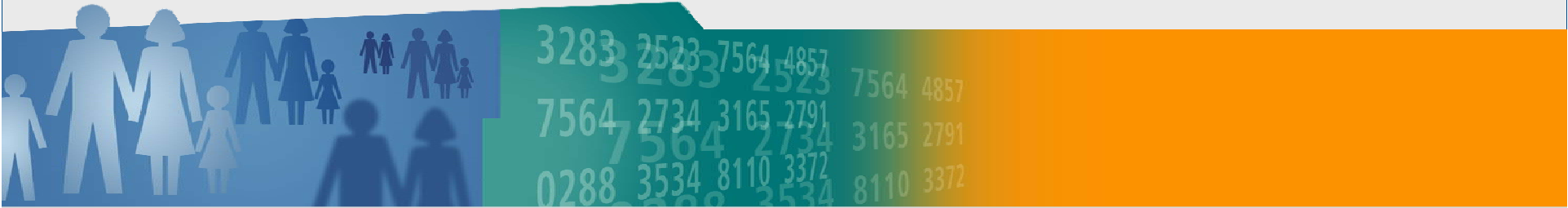


# Generalising Beyond Socioeconomic Status

- While the focus has been imputing SES onto administrative records, the methodology is more general
- So long as an empirical distribution is known and there are matching variables on both datasets, conceivably any variable could be imputed onto any other data source
- The key caveat, of course, is the importance of the independence assumption and the extent to which any induced bias can be quantified

# Acknowledgements

- Australian Research Council Discovery grant (DP0881616)
- Australian Research Council Strategic Partnership with Industry Research and Training (SPIRT) grant (C00107794)



[www.natsem.canberra.edu.au](http://www.natsem.canberra.edu.au)