A new method for confidentialising Census tables

Bruce Fraser
Data Access and Confidentiality
Methodology Unit
ABS

CENSUS ANALYSIS CONFERENCE 06
MAKING THE MOST OF YOUR CENSUS

Why confidentialise tables?

- Census and Statistics Act: (Results) shall not be published or disseminated in a manner that is likely to enable the identification of a particular person or organization
- Quality of output: Strong confidentiality protection leads to strong public trust and strong cooperation with the Census

Current procedures

- May withhold large tables for small subpopulations
- Introduce random error for very small cells
- Re-derive table totals and sub-totals
- These techniques allow the ABS to release detailed tables that would otherwise not be possible

Constraints of the current procedures

- Census output must be limited to a relatively small set of pre-determined categories
- Tables are inconsistent
- Random error carries through to the table totals / sub-totals

CENSUS ANALYSIS CONFERENCE 06
MAKING THE MOST OF YOUR CENSUS

- Inconsistency - for example different tables will give different values for the total population of a small area

Drivers for change

- Desire to give much greater flexibility to Census users in output
- Improvements in IT increase the risk of a confidentiality beach
  - We need more sophisticated methods
- Increased public scrutiny
  - We need defensible methods

CENSUS ANALYSIS CONFERENCE 06
MAKING THE MOST OF YOUR CENSUS

- Flexibility - in particular, to allow users to define their own geographies as aggregations of small area mesh blocks

New system

- Not yet signed-off
- Table cell have a non-zero probability that random error will be introduced (except for "key" statistics)
- There is much greater consistency than previously
- Protects against "differencing"

CENSUS ANALYSIS CONFERENCE 06
MAKING THE MOST OF YOUR CENSUS

- Differencing - deriving a small cell by subtracting one large cell from another one

Measuring the impact of the new system

- Measurement of the impact has been inspired by the χ^2 test of association
- Calculate $(o-e)^2/e$ for each interior cell
- Sum over rows and columns
- Rank the size of row and column totals
- Measure difference in grand total and in ranks

Table of $(o-e)^2/e$ values

	Employed full-time	Employed part-time	Unemployed full-time	Unemployed part-time	Not in the Labour Force	**Total**	*Rank*
15-24	2,139	753	73	1,282	151	**4,399**	*2*
25-44	1,131	68	15	4	681	**1,900**	*4*
45-64	920	33	2	31	1,152	**2,138**	*3*
65+	4,711	3,096	187	195	14,507	**22,695**	*1*
Total	**8,902**	**3,951**	**277**	**1,511**	**16,491**	**31,131**	
Rank	*2*	*3*	*5*	*4*	*1*		

CENSUS ANALYSIS CONFERENCE 06
MAKING THE MOST OF YOUR CENSUS

- 31,131 is the value of the usual chi-squared statistic
- Calculate the percentage change in this statistic
- Also look at changes in the row and column ranks

Impact at Australia level

Region	Population	Table	Protection	Change in chi²	Maximum change in row rank	Maximum change in column rank
Australia	18,769,249	CoB x YoA	New	-0.0003%	0	0
			Old	0	0	0
		Age x LF (Males)	New	-0.0006%	0	0
			Old	0	0	0
		Age x LF (Females)	New	0.0006%	0	0
			Old	0	0	0
		Ind x Occ	New	-0.0013%	0	0
			Old	0	0	0

CENSUS ANALYSIS CONFERENCE 06
MAKING THE MOST OF YOUR CENSUS

- Very small level of noise added to large counts
- There is a measurable impact but it is negligible

Impact for ACT

Region	Population	Table	Protection	Change in chi²	Maximum change in row rank	Maximum change in column rank
ACT	309,998	CoB x YoA	New	-0.03%	1	0
			Old	0.19%	0	0
		Age x LF (Males)	New	0.04%	0	0
			Old	0.00	0	0
		Age x LF (Females)	New	-0.03%	0	0
			Old	-0.01%	0	0
		Ind x Occ	New	0.12%	0	0
			Old	0.07%	0	0

CENSUS ANALYSIS CONFERENCE 06
MAKING THE MOST OF YOUR CENSUS

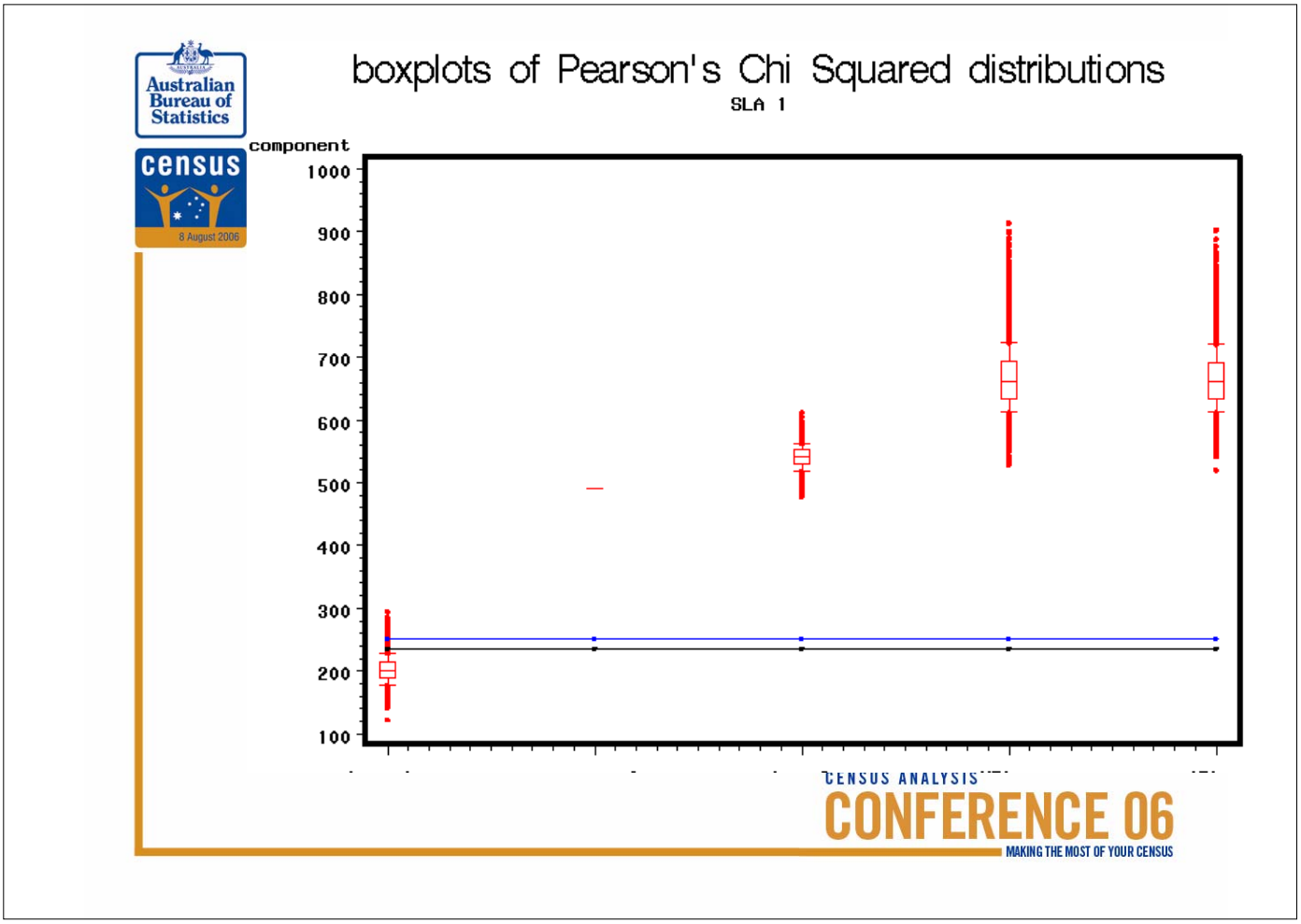
- Impacts are more noticable for a moderately large population of 300,000
- Impacts are negligible

Impact for Buloke Shire

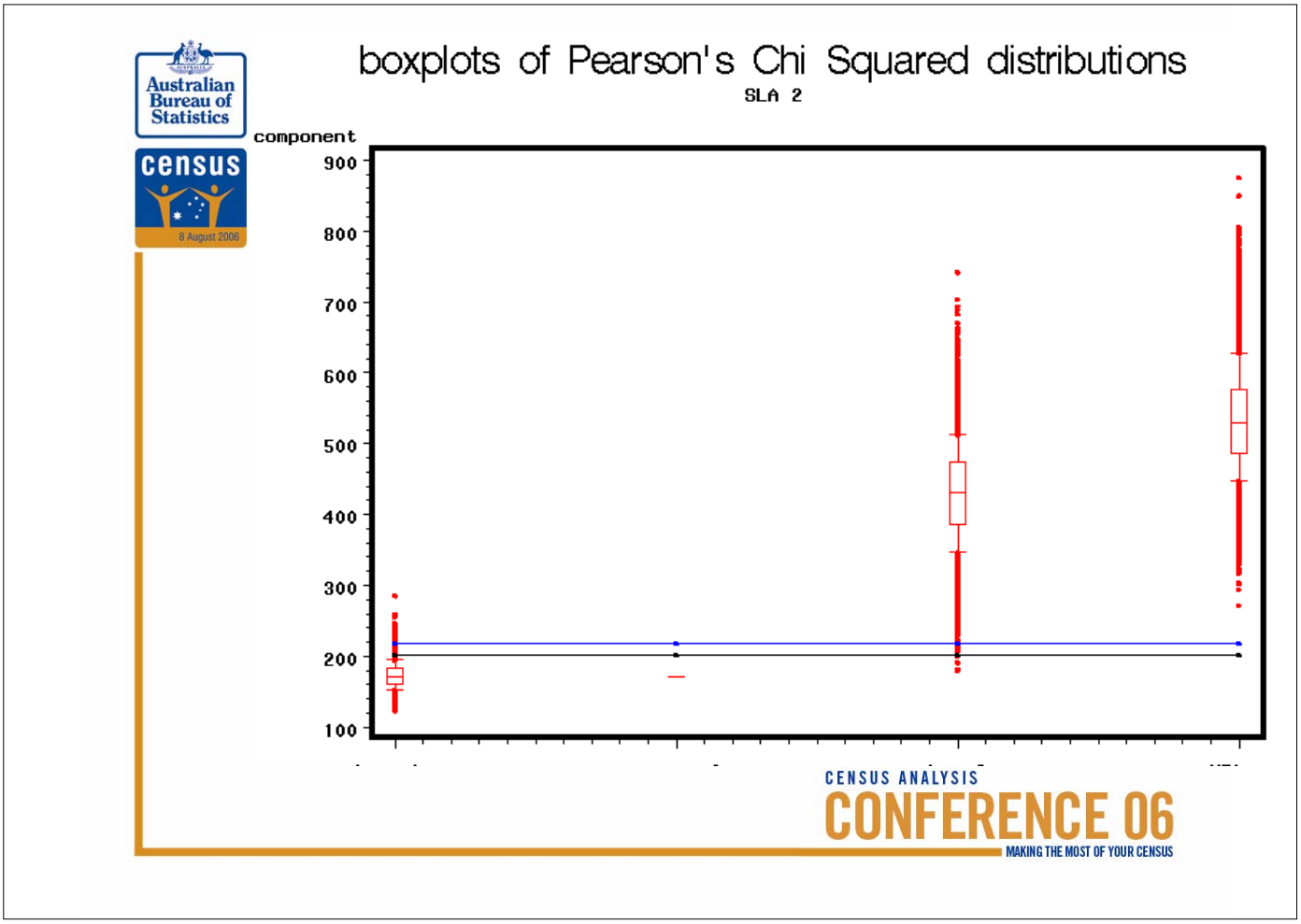
Region	Population	Table	Protection	Change in chi²	Maximum change in row rank	Maximum change in column rank
Buloke LGA	7,147	Age x LF (Males)	New	4.92%	2	1
			Old	1.34%	0	1
		Age x LF (Females)	New	3.80%	1	0
			Old	-1.89%	1	0
		Ind x Occ	New	0.02%	5	1
			Old	-0.10%	2	1

CENSUS ANALYSIS CONFERENCE 06
MAKING THE MOST OF YOUR CENSUS

- Impacts are more noticeable for populations of approximately 10,000, although still not large enough to distort conclusions



- SLA with population 1200
- chi-squared statistic is biased upwards by both old and new method
- bias is greater with new method
- variability is also greater
- but this work will help us describe this effect to users, and so help them to compensate



- SLA2 has population 100
- less difference between old and new methods
- both methods bias the chi-squared statistic upwards

Conclusion

- New method has a similar impact to the current method for small to moderately sized populations
- New method has a negligible impact for large populations
- New method improves consistency of estimates within regions, and allows for a greater range of output
- We should help to compensate for biases introduced by confidentiality protection