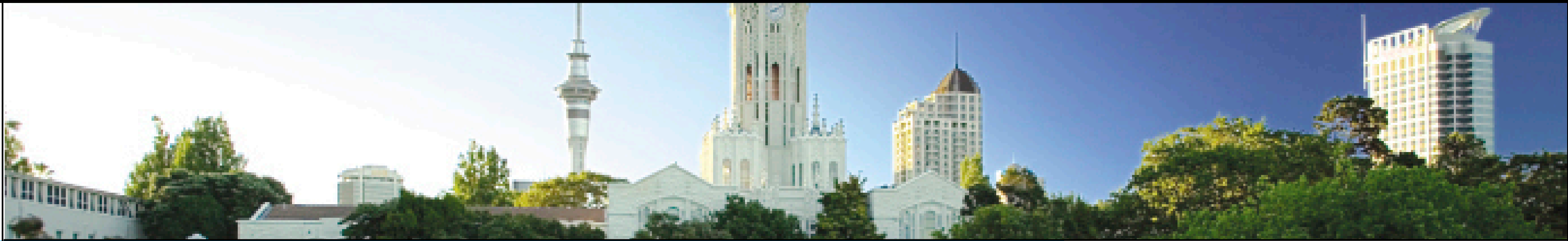


Creating synthetic data using composites of similar individuals

COMPASS Colloquium
August 2013

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

Barry Milne
COMPASS Research Centre
University of Auckland
New Zealand
www.compass.auckland.ac.nz

MINISTRY OF BUSINESS, INNOVATION & EMPLOYMENT
HIKINA WHAKATUTUKI

DISCLAIMER: *Access to the data used in this study was provided by Statistics New Zealand under conditions designed to give effect to the security and confidentiality provisions of the Statistics Act 1975. The results presented in this study are the work of the author, not Statistics New Zealand.*

Overview

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

1. **Approach and Methods** - how we created synthetic data using 'composite clusters'

2. **Quality** – how well did our synthetic data reproduce the "real" data in Census 2006

3. **Confidentiality** – how well did our synthetic data resist attempts to reveal "real" individuals in Census 2006

Why synthetic data?

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

- Need representative birth cohort for simulation model, and need it released (for wide use)
  - 2006 Census is representative and has many of the variables we use to get our model running
  - 2006 Census micro-data unable to be released, but what if we had something that looked like 2006 Census micro-data but didn't contain any actual individual….
- Achievable if create 'synthetic birth cohort'
  - Usual approaches are perturbation or multiple imputation, but we're trying something different

Composite clusters

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

- Creating a synthetic base file of composite individuals

Methods

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

- Subset 2006 Census to include just new-borns (0-year olds) and their parents
  - Randomly select 10,000 (Processing speed)
- Calculate distance between each of the 10,000, based on 2006 Census characteristics.
  - Done separately for two-parent families, single mum families and single dad families (for consistency)
- Choose the closest 2 ranks to form 10,000 clusters of 3 individuals

Methods

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

- Randomly choose, characteristic-by-characteristic, which member of the cluster's characteristics contributes to the synthetic individual

- Voilà! A synthetic basefile of 10,000 composite individuals

Questions?

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

The University of Auckland New Zealand

Quality

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

The University of Auckland New Zealand

Quality

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

- Synthetic data should faithfully represent distributions and inter-relations of real data

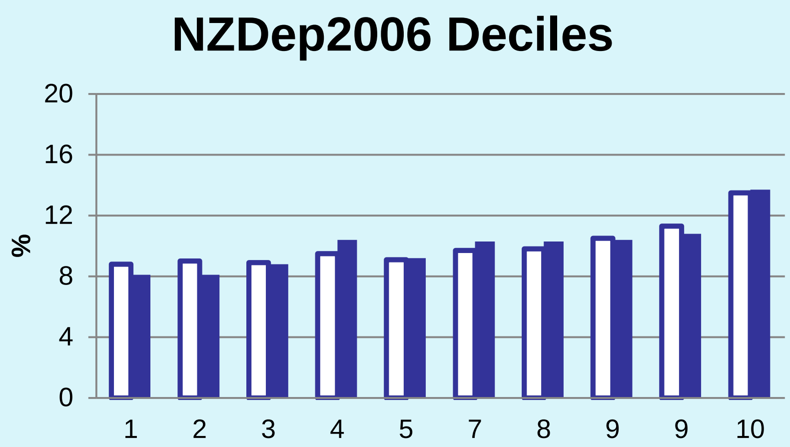
- Distributions
  - Proportions in groups
  - Mean, SD & shape of continuous variables

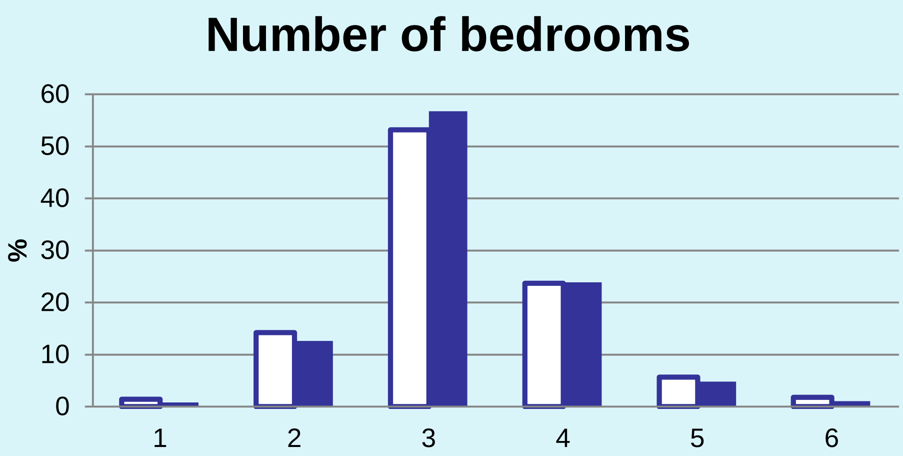
- Inter-relations
  - Variables strongly correlated in the real data should also be strongly correlated in the synthetic data

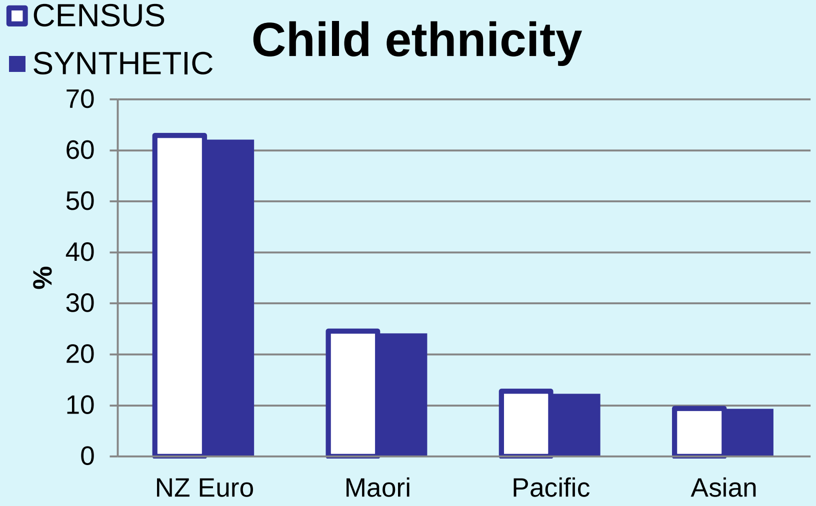
The University of Auckland New Zealand

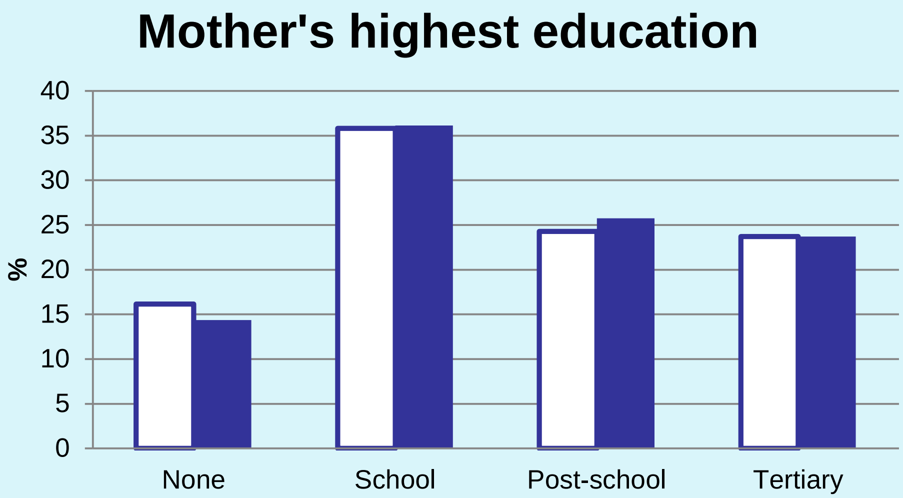
Quality
- Distributions

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau









Quality
- Distributions

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

		Mean	SD	Percentiles				
				10	**25**	**50**	**75**	**90**
Mum's age	**Census**	30.7	6.7	22	26	31	35	38
	Synth	30.5	6.2	22	26	31	34	37
Dad's age	**Census**	33.9	7.0	25	30	34	38	42
	Synth	33.6	6.2	26	30	34	37	41
Years at address	**Census**	2.96	4.40	0	0	2	4	7
	Synth	2.74	3.85	0	0	2	4	6

Quality
- Inter-relations

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

		Var1	Var 2	Var 3	Var 4	……….	Var n
Var 1	Census	1	.12	-.23	.34	……….	.45
Var 2	Census		1	.02	-.13	……….	.24
Var 3	Census			1	.42	……….	-.01
Var 4	Census				1	……….	.17
. .	Census					1	……….
Var n	Census						1

Quality
- Inter-relations

		Var1	Var 2	Var 3	Var 4	……….	Var n
Var 1	Census **Synthetic**	1 **1**	.12 **.05**	-.23 **-.29**	.34 **.21**	………. **……….**	.45 **.51**
Var 2	Census **Synthetic**		1 **1**	.02 **-.04**	-.13 **-.21**	………. **……….**	.24 **.19**
Var 3	Census **Synthetic**			1 **1**	.42 **.35**	………. **……….**	-.01 **.06**
Var 4	Census **Synthetic**				1 **1**	………. **……….**	.17 **.05**
. .	Census **Synthetic**					1 **1**	………. **……….**
Var n	Census **Synthetic**						1 **1**

Quality
- Inter-relations

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

		Var1	Var 2	Var 3	Var 4	....	Var n
Var 1	Census **Synthetic**	1 **1**	.12 **.05**	-.23 **-.29**	.34 **.21**	.... **..........**	.45 **.51**
Var 2	Census **Synthetic**		1 **1**	.02 **-.04**	-.13 **-.21**	.... **..........**	.24 **.19**
Var 3	Census **Synthetic**			1 **1**	.42 **.35**	.... **..........**	-.01 **.06**
Var 4	Census **Synthetic**				1 **1**	.... **..........**	.17 **.05**
. .	Census **Synthetic**					1 **1**	.... **..........**
Var n	Census **Synthetic**						1 **1**

Quality
- Inter-relations

- Correlation between
  - two-way correlations among Census variables
  - two-way correlations among Synthetic variables
  - **r=0.66** (n=1596 pairwise correlations for 57 vars)
- Moreover, associations aren't dampened. Mean magnitude of correlations
  - CENSUS: r = .097
  - SYNTHETIC: r = .102
- Correlations in Census tend to be replicated in synthetic file
  - Suggests inter-relations have been maintained

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

The University of Auckland New Zealand

Confidentiality

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

Confidentiality

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

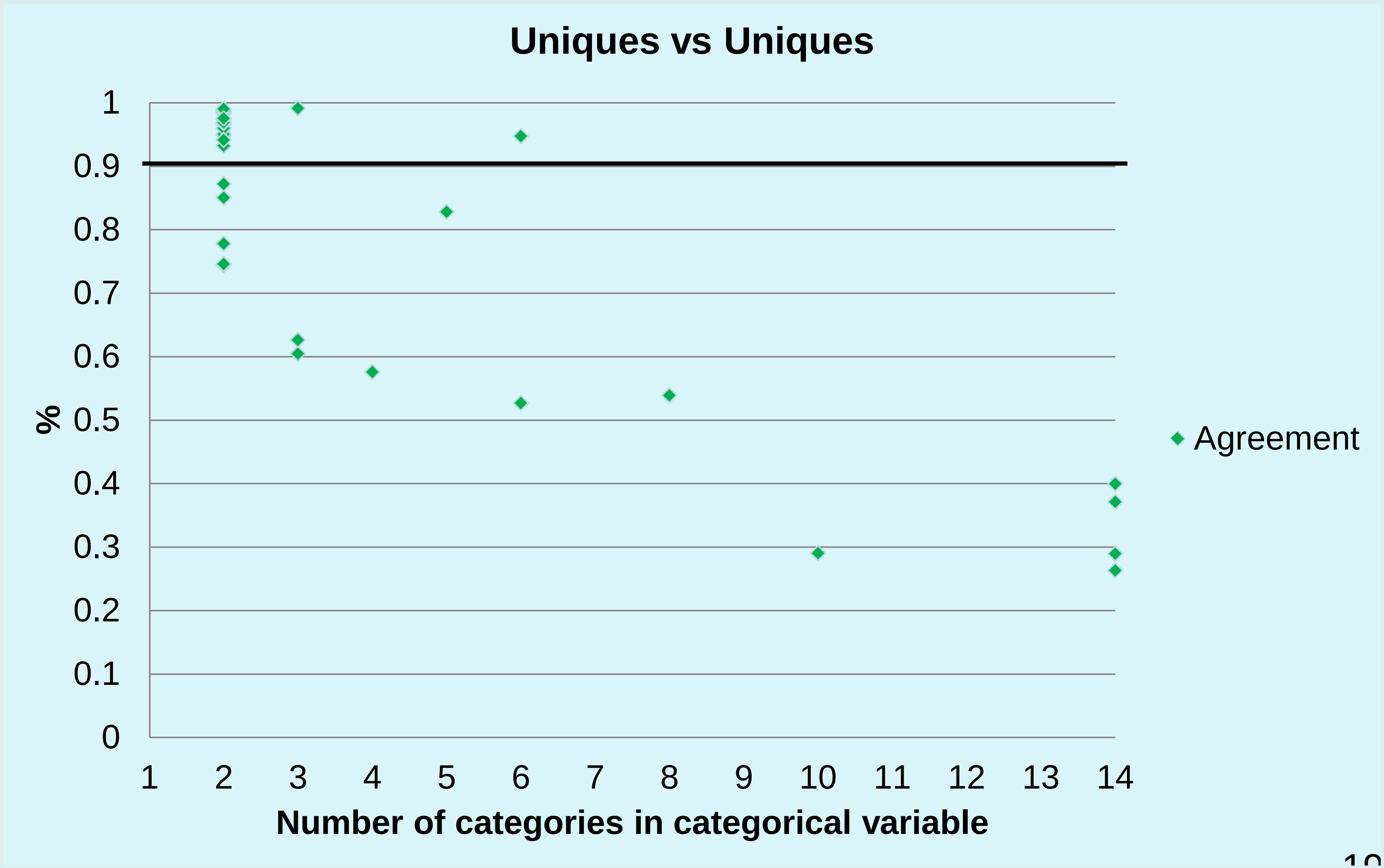
- ‘Hacker scenarios’
- Could a ‘hacker’ gaining access to our synthetic file learn anything ‘new’ about a real individual (about whom they had some basic information).
- Process
  - Find ‘uniques’ in the synthetic data using ‘strongly identifying’ information (ethnicity [M/F/C], age [M/F], sex [C])
  - Are there individuals with the same characteristics in the 2006 Census? If so, can we learn ‘sensitive’ information about these real individuals based on their synthetic data?

Confidentiality - Uniques

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

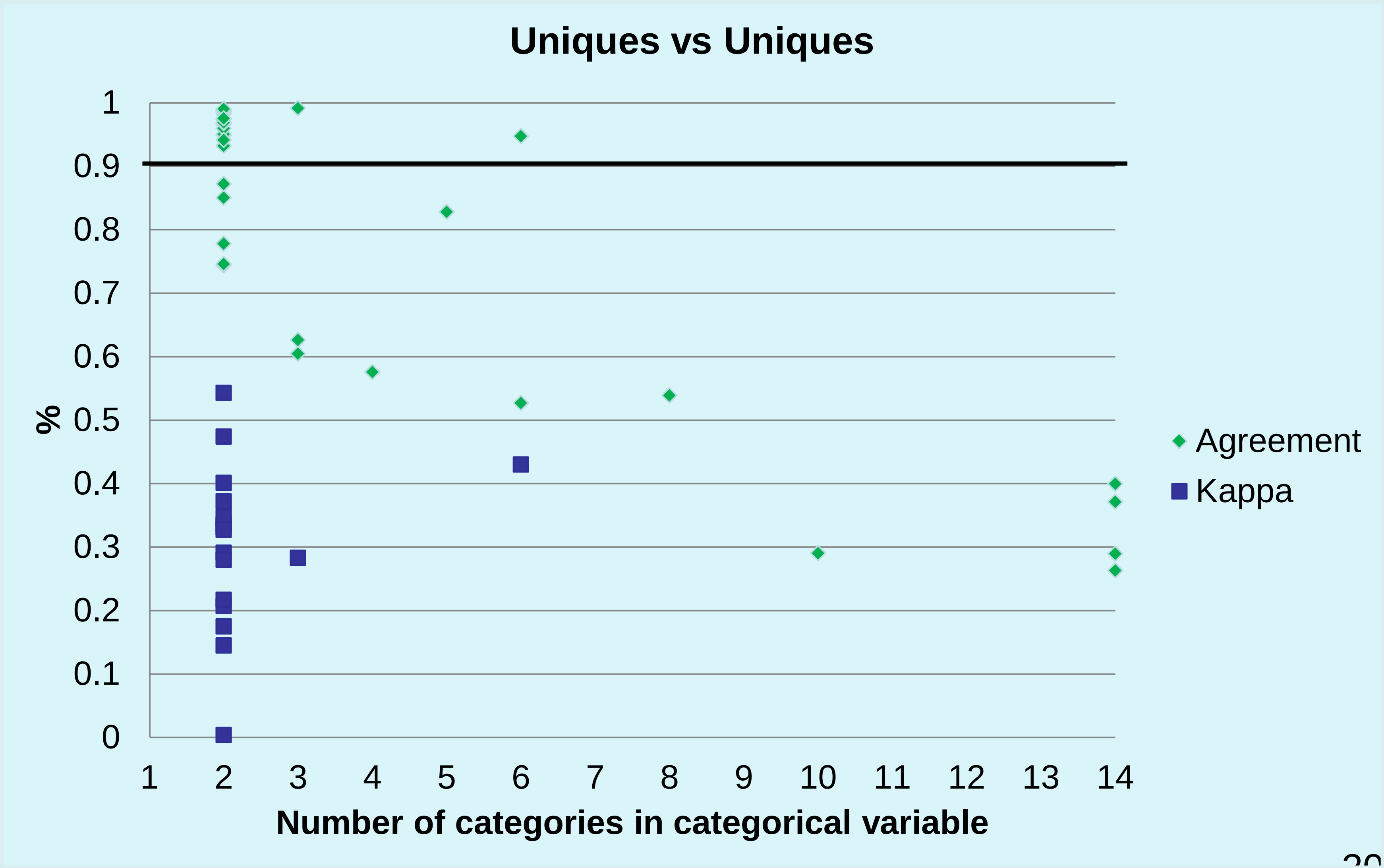
- Of synthetic individuals with ‘unique’ characteristics (C sex, M, F & C ethnicity, M & F age; 48.6%)
  - 62.5% don’t exist in the Census
  - 13.3% are unique in the Census
  - 24.1% are shared by 2+ people in the Census
- What is the level of agreement for non-identifying characteristics?
  - Synthetic uniques vs. <u>unique</u> counterpart in Census
  - Synthetic uniques vs. <u>non-unique</u> counterparts in Census
  - Not allowed to exceed 90%

Confidentiality
- Agreement with real data



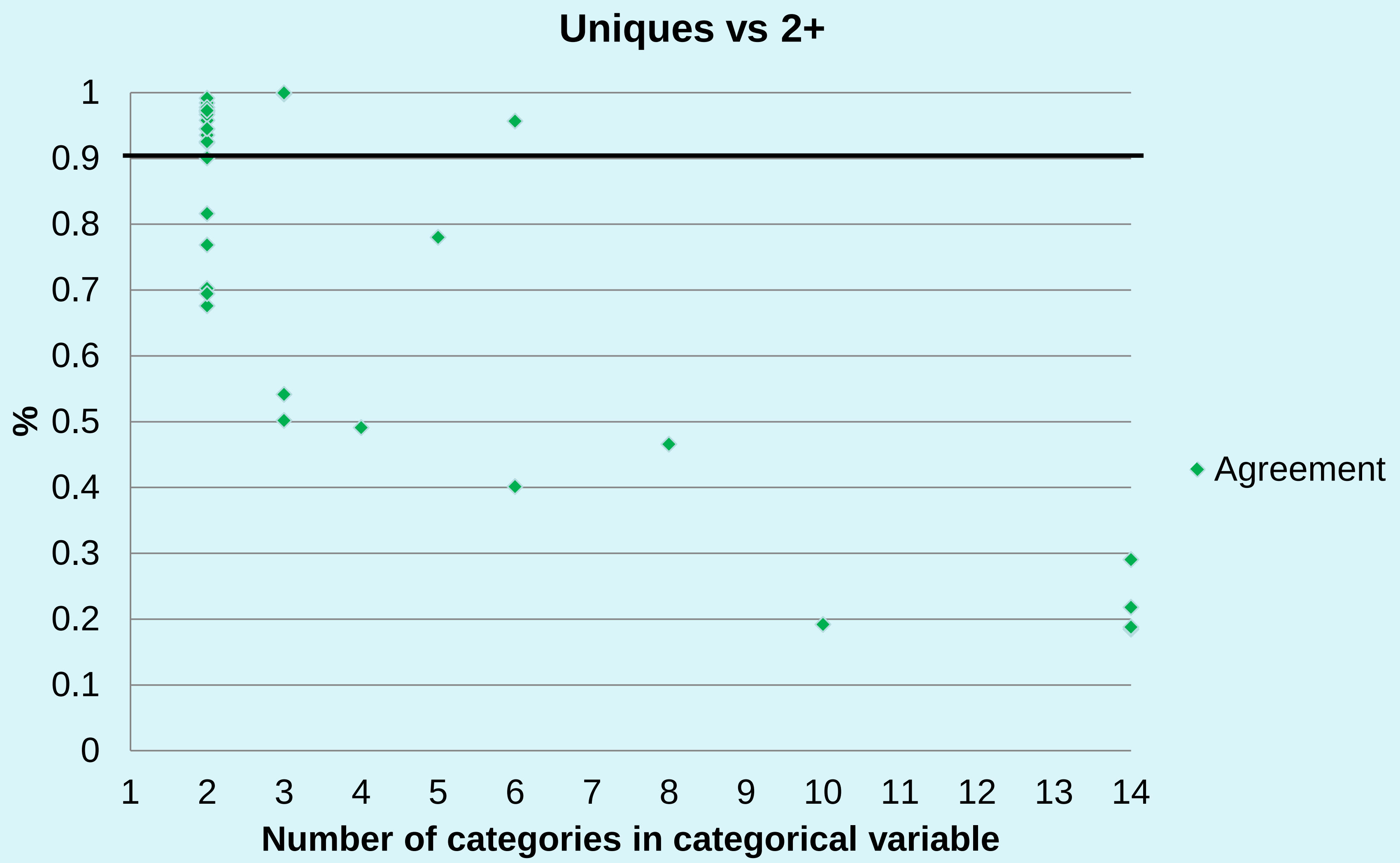
Confidentiality
- Agreement with real data

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau



Confidentiality
- Agreement with real data

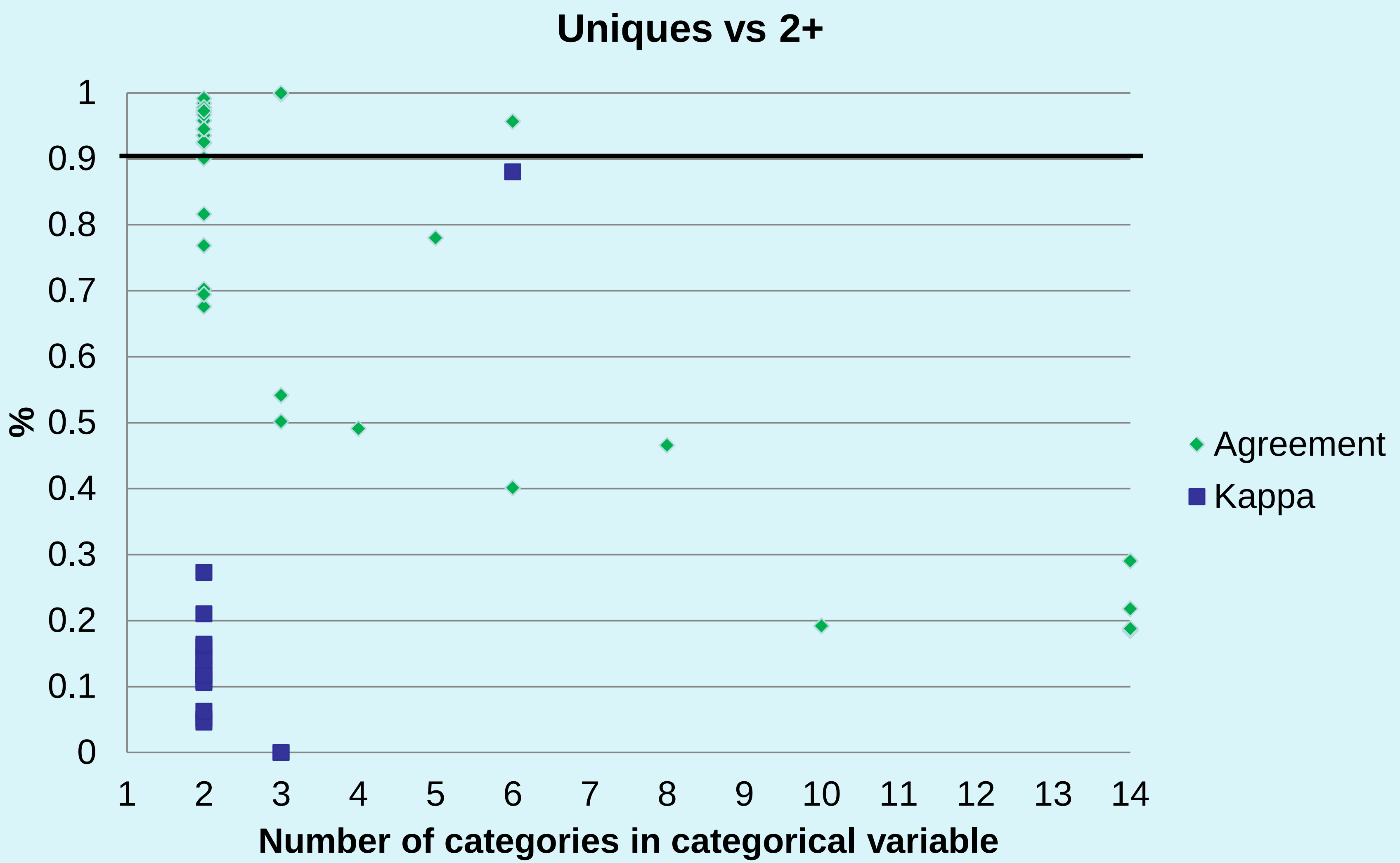
COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau



The University of Auckland New Zealand

Confidentiality
- Agreement with real data

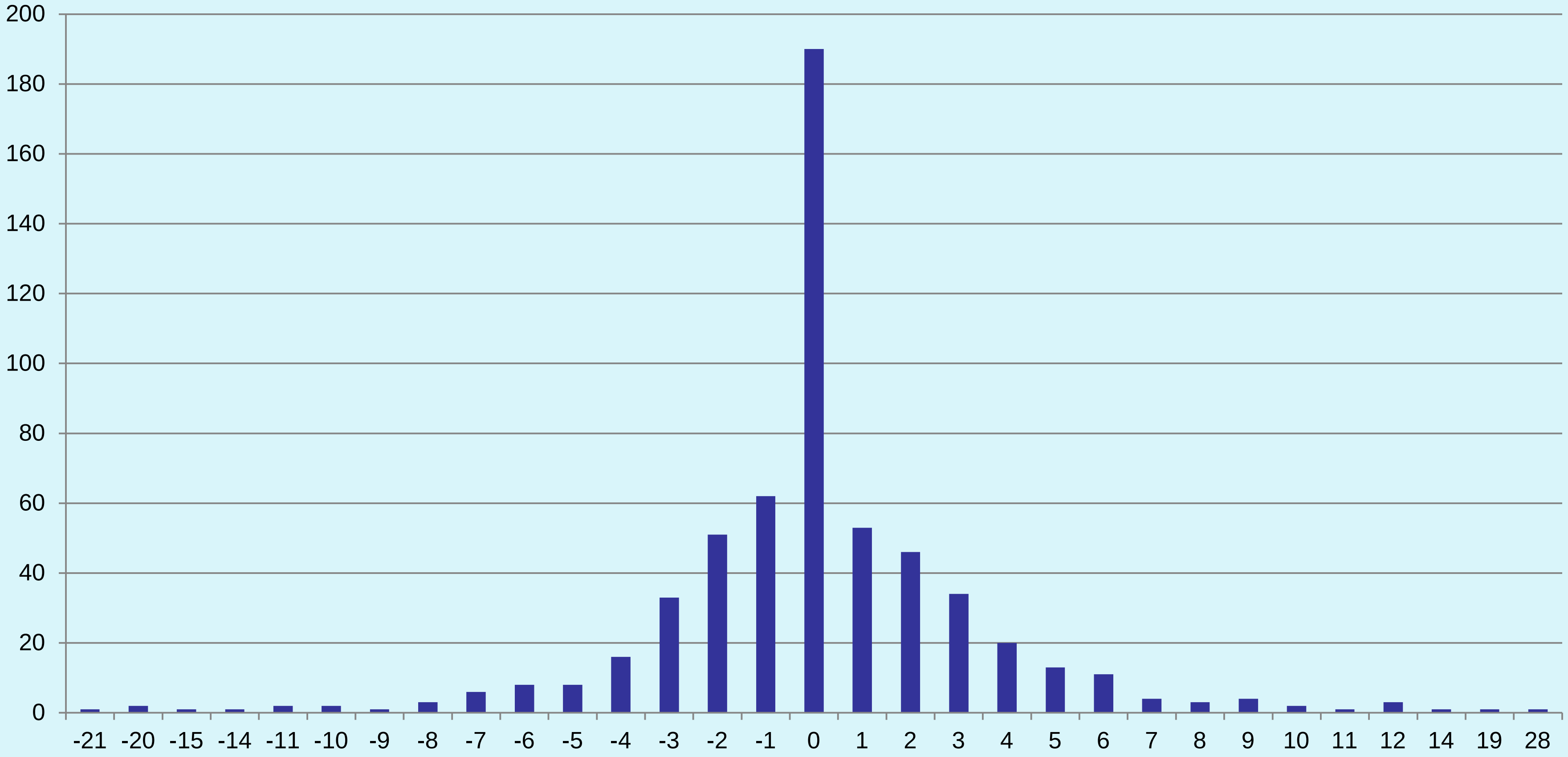
COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau



The University of Auckland
New Zealand

Results
- Confidentiality

- Years at address: 32% perfect agreement



Results
- Confidentiality

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

- ‘On-diagonals’ never exceed 90% EXCEPT for very low base rate characteristics
  - High probability of hitting on-diagonals by chance
- For most exceeding 90% kappa suggests far from perfect agreement, except
  - child_depend_family_type_code (6 categories)
    - Couple with dependent child
    - Couple with dependent child & adult
    - Couple with dependent child & unknown
    - One parent with dependent child
    - One parent with dependent child & adult
    - One parent with dependent child & unknown

Conclusions

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

Conclusions

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

1. Creating synthetic data using ‘composite clusters’ is achievable and (relatively) quick

2. Data is high quality
   - Distributions closely match those of Census
   - Inter-relations approximate those of Census (both in directionality & magnitude)

3. Data meets confidential requirements
   - Small overlap between ‘uniques’ in synthetic file and ‘uniques’ in Census; and ‘uniques’ don’t reliably reveal anything factual about a ‘real’ individual

Conclusions

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

- How does it compare with multiple imputation?
  - Direct test is underway
- Composite approach potentially suitable to any synthetic data creation
  - Processing power may be issue
- Flexibility to adequately balance quality and confidentiality
  - Quality poor? Use fewer matches
  - Confidentiality compromised? Use more matches

Questions?

COMPASS RESEARCH CENTRE
FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND
Whare Wānanga o Tāmaki Makaurau

The University of Auckland New Zealand