Balancing Access and Confidentiality: The Role of Synthetic Data in Official Statistics -- Presentation 3: New Zealand

Alistair Ramsden, Senior Design Analyst,
Methods & Design, Stats NZ -- Sept-2024

Overview

- What is statistical confidentiality?
- What is synthetic data?

- What are the benefits of synthetic data?
- How can we measure the safety and utility of synthetic data?

- Methods for use of synthetic data at Stats NZ
- Examples of use of synthetic data at Stats NZ
- Plans for future use of synthetic data at Stats NZ

What is statistical confidentiality?

"So what exactly is statistical confidentiality?

Simply put, it is the stewardship of data to be used for statistical purposes.

Stewardship, as expressed in statistical confidentiality, is an active embrace of responsibility for *both protecting data and ensuring its beneficial use*.

Explicitly it requires *proper practices for both providing and restricting access* to data products."

from 'Statistical Confidentiality Principles and Practice'

(Duncan G, Elliot M, and Salazar-González J-J, 2011)

What is synthetic data?

Data which mimics original data, provides privacy guarantees

(Machanavajjhala, Kifer, Abowd, Gehrke, and Vilhuber, 2008)

Model generated data, not real data

(OECD, 2007)

Production data not obtained by direct measurement

(McGraw-Hill, 2003)

Inferentially valid, confidentialised, public-use microdata

(Rubin, 1993)

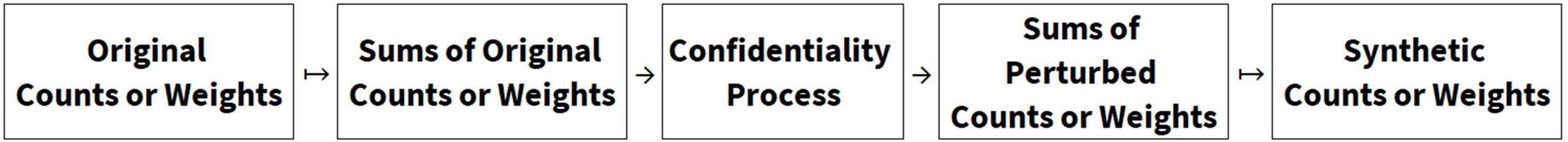
What is synthetic data?

- Synthetic data is any *at least partially* noised or perturbed data, with or without some retained *inferential validity*

- Any confidentialised descriptive statistics summary table can be reverse-transformed into a synthetic data unit record set

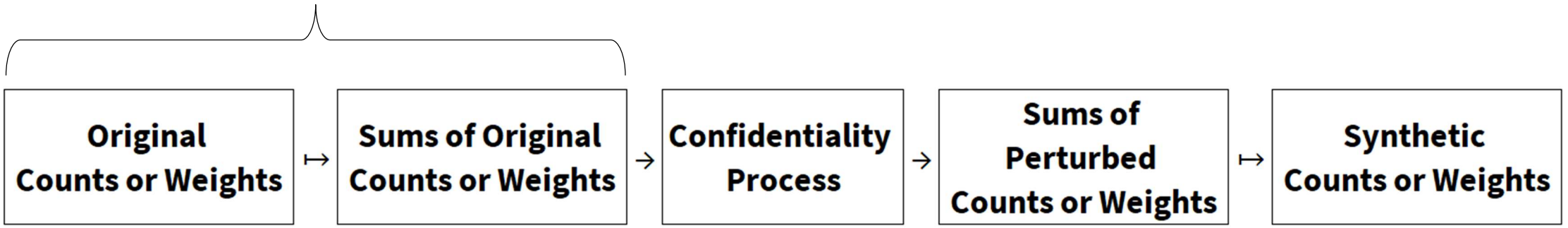
- Any inferentially valid results in synthetic data are substantively identical to results in the original data – and providing customers with metadata about retained inferential validity is essential

What is synthetic data?



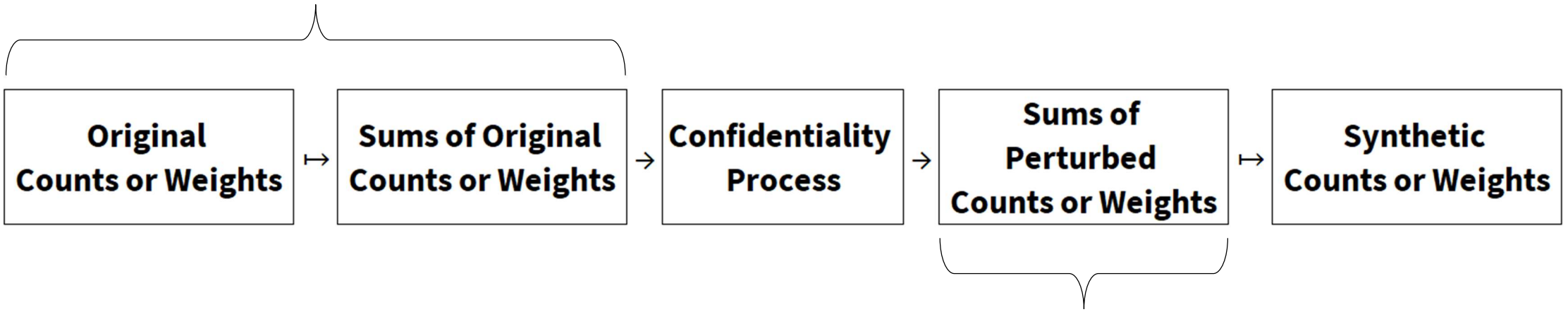
What is synthetic data?

This is unconfidentialised data securely held by NSOs



What is synthetic data?

This is unconfidentialised data securely held by NSOs



This is confidentialised, publishable data

What is synthetic data?

X Unit Records	Variable A	Variable B	...	Variable N	Original Counts or Weights
1	value	value	...	value	w1
2	value	value	...	value	w2
3	value	value	...	value	w3
...	...	...	...	...	...
X	value	value	...	value	wX

↦

Variable A	Variable B	...	Variable N	Y Sums of Original Counts or Weights
value A(1)	value B(1)	...	value N(1)	s1
value A(1)	value B(1)	...	value N(2)	s2
value A(1)	value B(1)	...	...	
value A(1)	value B(1)	...	value N(n(N))	s4
value A(1)	value B(2)	...	value N(1)	s5
value A(1)	value B(2)	...	value N(2)	s6
value A(1)	value B(2)	...	...	s7
value A(1)	value B(2)	...	value N(n(N))	s8
...	...	...	...	...
value A(n(A))	value B(n(B))	...	value N(n(N))	sY

↓

Confidentiality Process

↓

X' Unit Records	Variable A	Variable B	...	Variable N	Synthetic Counts or Weights
1'	value'	value'	...	value'	w1'
2'	value'	value'	...	value'	w2'
3'	value'	value'	...	value'	w3'
...	...	...	...	...	...
X'	value'	value'	...	value'	wX'

↤

Variable A	Variable B	...	Variable N	Y' Sums of Perturbed Counts or Weights
value A(1)	value B(1)	...	value N(1)	s1'
value A(1)	value B(1)	...	value N(2)	s2'
value A(1)	value B(1)	...	...	s3'
value A(1)	value B(1)	...	value N(n(N))	s4'
value A(1)	value B(2)	...	value N(1)	s5'
value A(1)	value B(2)	...	value N(2)	s6'
value A(1)	value B(2)	...	...	s7'
value A(1)	value B(2)	...	value N(n(N))	s8'
...	...	...	...	...
value A(n(A))	value B(n(B))	...	value N(n(N))	sY'

What is synthetic data? A synthetic example

X Unit Records	Variable A	Variable B	...	Variable N	Original Counts or Weights
1	Small	Red	...	Triangle	1
2	Small	Red	...	Square	1
3	Large	Orange	...	Square	1
...	...	...	...	...	...
X	Large	Blue	...	Triangle	1

↦

Variable A	Variable B	...	Variable N	Y Sums of Original Counts or Weights
Small	Red	...	Triangle	1
Small	Red	...	Square	2
Small	Red	...	...	3
Small	Red	...	Circle	4
Small	Orange	...	Triangle	5
Small	Orange	...	Square	6
Small	Orange	...	...	7
Small	Orange	...	Circle	8
...	...	...	...	9
Large	Blue	...	Circle	10

↓

Confidentiality Process

↓

X' Unit Records	Variable A	Variable B	...	Variable N	Synthetic Counts or Weights
1'	Small	Red	...	Square	1
2'	Small	Red	...	Square	1
3'	Small	Red	...	Square	1
...	...	...	...	...	...
X'	Large	Blue	...	Circle	1

↤

Variable A	Variable B	...	Variable N	Y' Sums of Perturbed Counts or Weights
Small	Red	...	Triangle	0
Small	Red	...	Square	3
Small	Red	...	...	3
Small	Red	...	Circle	3
Small	Orange	...	Triangle	6
Small	Orange	...	Square	6
Small	Orange	...	...	6
Small	Orange	...	Circle	9
...	...	...	...	9
Large	Blue	...	Circle	9

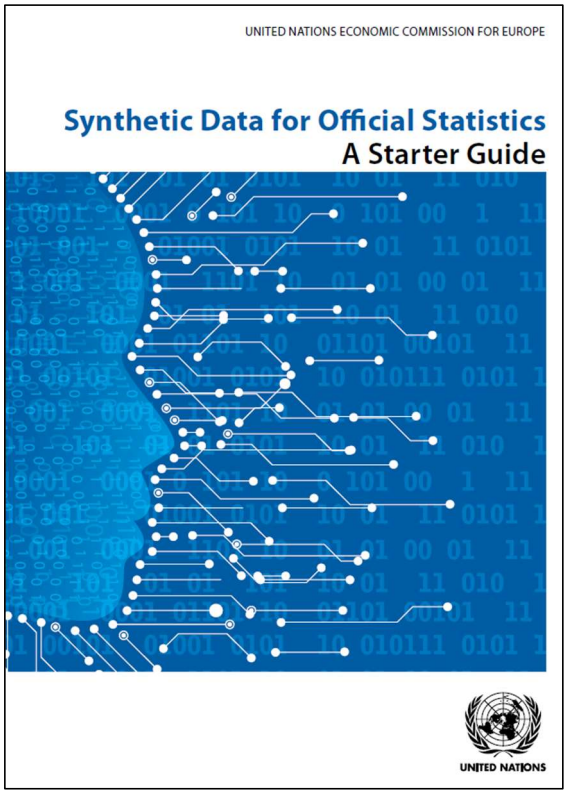
What are the benefits of synthetic data?

- Synthetic data, with or without some retained inferential validity, *can be used for testing, and training, and can improve data safety and data security* while doing so

- Data with *appropriate* statistical disclosure control / uncertainty, yet retaining *some* inferential validity, can be safely published

- *Providing users with metadata* about how data was synthesised / confidentialised, and/or what inferential validity is retained in the data, is essential

How can we measure the safety and utility of synthetic data?



Chapter 4: Disclosure considerations for synthetic data 33
4.1 Measures of disclosure risk for synthetic data 34
4.2 Privacy-preserving techniques 38
4.3 Other methods for disclosure evaluation 56
4.4 Tips to get started 60

Chapter 5: Utility measures for evaluating synthetic data 63
5.1 Specific utility measures 65
5.2 General utility measures giving a single measure 68
5.3 Methods to explore aspects of utility 72
5.4 Tips for getting started 78

https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf

Figure 10 Decision tree to aid in disclosure control considerations

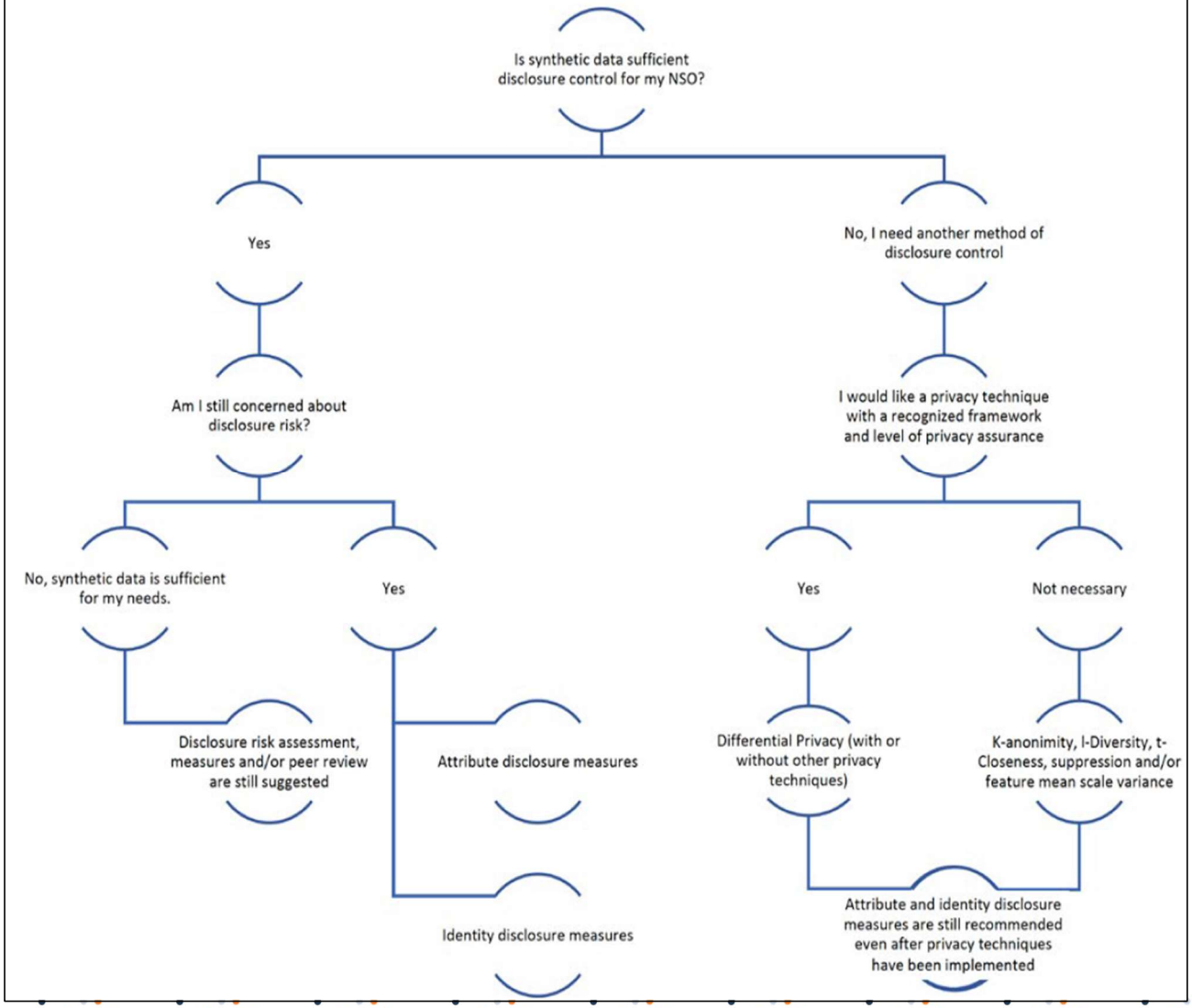
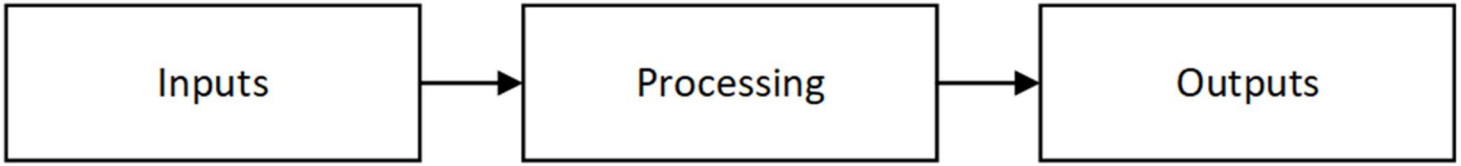


Table 20 Summary of utility measures

Method	Measure or procedure	Reference	Measure acronym
Specific	Confidence interval overlap	Karr et al. (2006)	
	Achieving specific outcomes	Kaloskampis et al. (2020), Jordon et al. (2018) and Slokom et al. (2021)	
General Single measure from propensity score	Propensity *score* mean squared error	Karr et al. (2006) and Woo et al. (2009)	pMSE
	Kolmogorov- Smirnov Statistic comparing propensity scores for original and synthetic data	Bowen et al. (2021)	SPECKS
	Other comparisons of propensity scores for original and synthetic data e.g., Wilcoxon signed rank statistic	Raab et al. (2021)	U
	Percentage over 50% of combined records correctly predicted by the propensity score	Raab et al. (2021)	PO50
General Single measure from tables	Voas-Williamson statistic	Voas and Williamson (2001)	VW
	Freeman-Tukey (FT)	Voas and Williamson (2001)	FR
	Likelihood ratio statistic from tables and other members of the divergence family	Voas and Williamson (2001)	G
	Jensen-Shannon Divergence	Fuglede and Topsoe (2004)	JSD
	Bhattacharyya metric	Bhattacharyya (1943)	dBhatt
	Mean absolute difference in densities	Raab et al. (2021)	MabsDD
	Difference of correlation matrices	Kaloskampis et al. 2019	
	Weighted mean absolute difference in densities	Raab et al. (2021)	WMabsDD
Methods for exploring utility	Comparing histograms by visualisation and summary measures	http://www.synthpop.org.uk and Kaloskampis et al. (2020)	
	Comparing cross-tabulations from marginal distributions	Raab (2011) and NIST (2021)	
	Comparing and visualising other summary statistics (e.g. Pearson Correlations)	Beaulieu-Jones et al. (2019) and Kaloskampis et al. (2020)	

Methods for use of synthetic data at Stats NZ

Synthetic data method – proof of concept general considerations



Is the proposed synthetic data product fit for purpose?

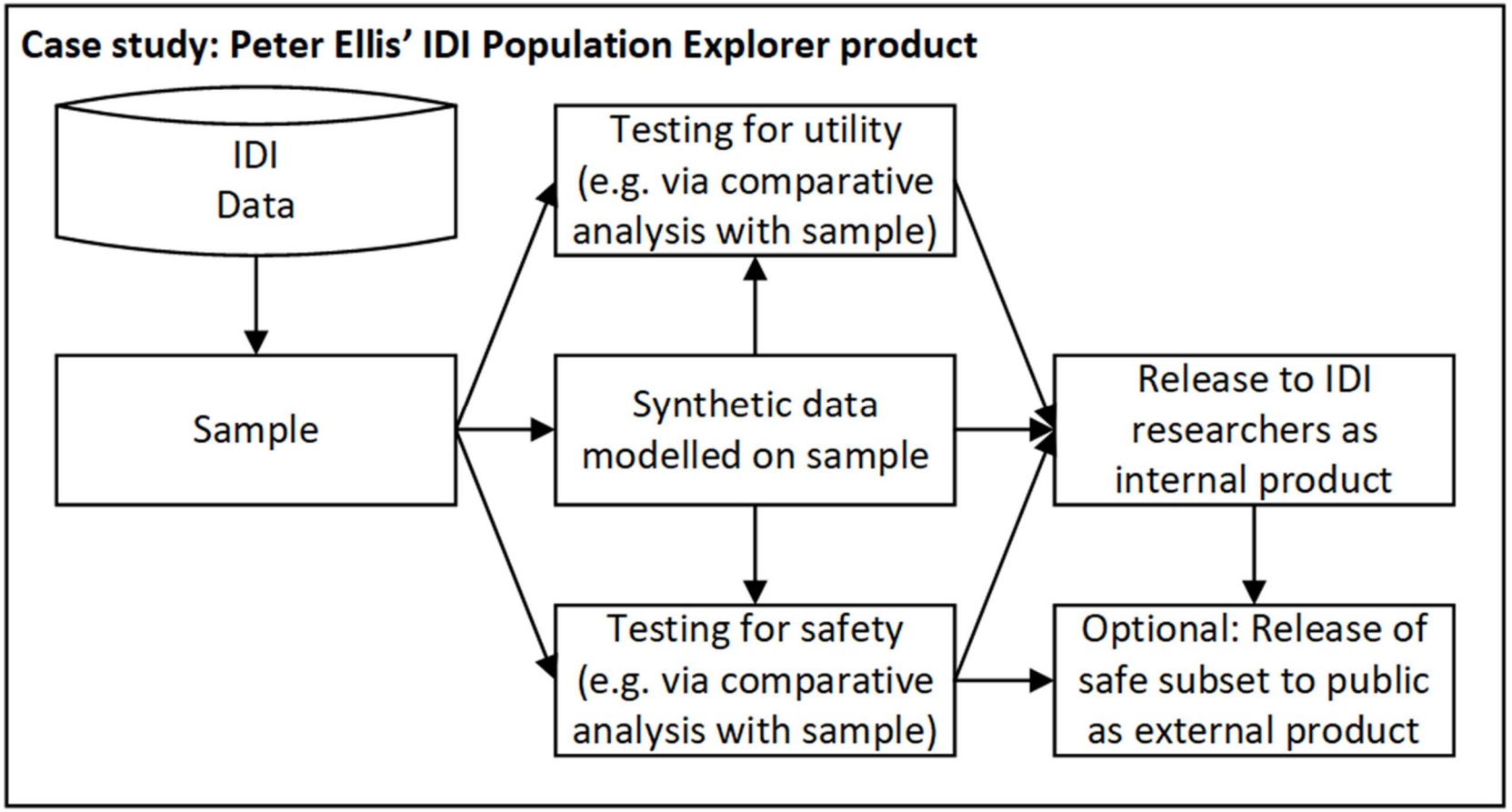
* What is the objective, and who are the users, of the product?

* Utility / accuracy of univariate and multivariate distributions of interest?

* Safety / inaccuracy of unit records?

* Does product align with IPSaC policy, relevant confidentiality standard, etc?

Methods for use of synthetic data at Stats NZ



Examples of use of synthetic data at Stats NZ

- 2018 Census and 2023 Census output tables; Integrated Data Infrastructure (IDI) output tables; as well as many other confidentialised descriptive statistics summary tables

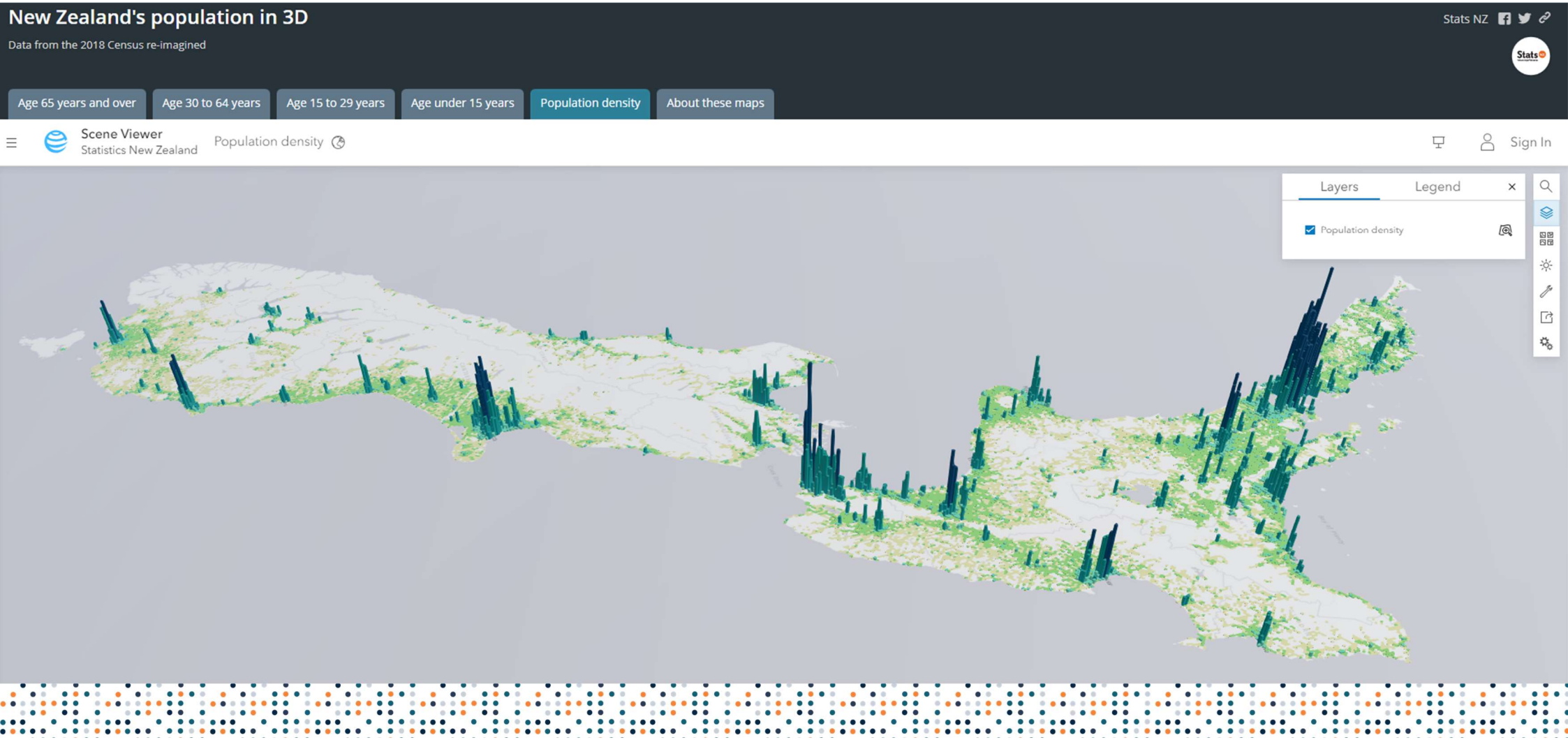
- A significant proportion of Stats NZ published data is *substantively identical* to synthetic data with some inferential validity retained

- Can be used as source data, e.g. for a Fully Conditional Specification (FCS) model, which can derive an informatively identical synthetic data set *like* the original data

Examples of use of synthetic data at Stats NZ

- 2023 New Zealand Grid Geography project
- 2021 *Nicholson Consulting / Kōtātā Insight* 'He Ara Poutama mō te reo Māori - Forecasting te reo Māori Speakers in Aotearoa, New Zealand' project
- Curated, limited variable synthetic data, based on 2013 Census data, for internal testing of 2018 Census output mode options
- 2008-2011 Synthetic Unit Record Files (SURFS) for Schools

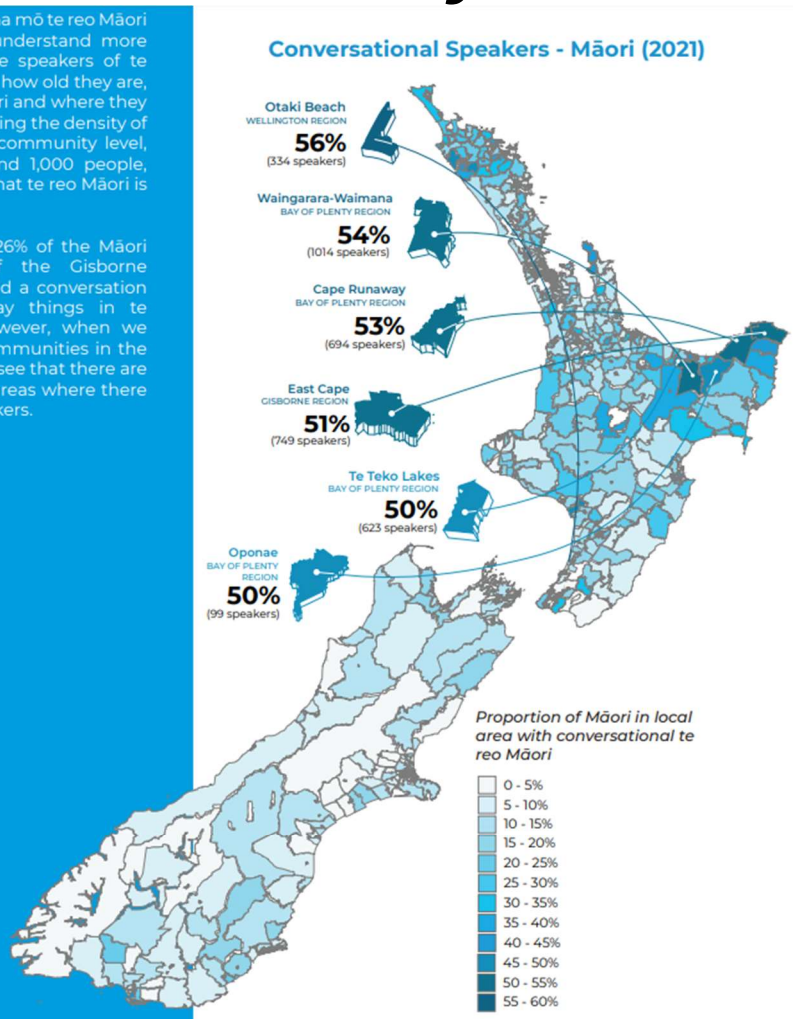
Examples of use of synthetic data at Stats NZ

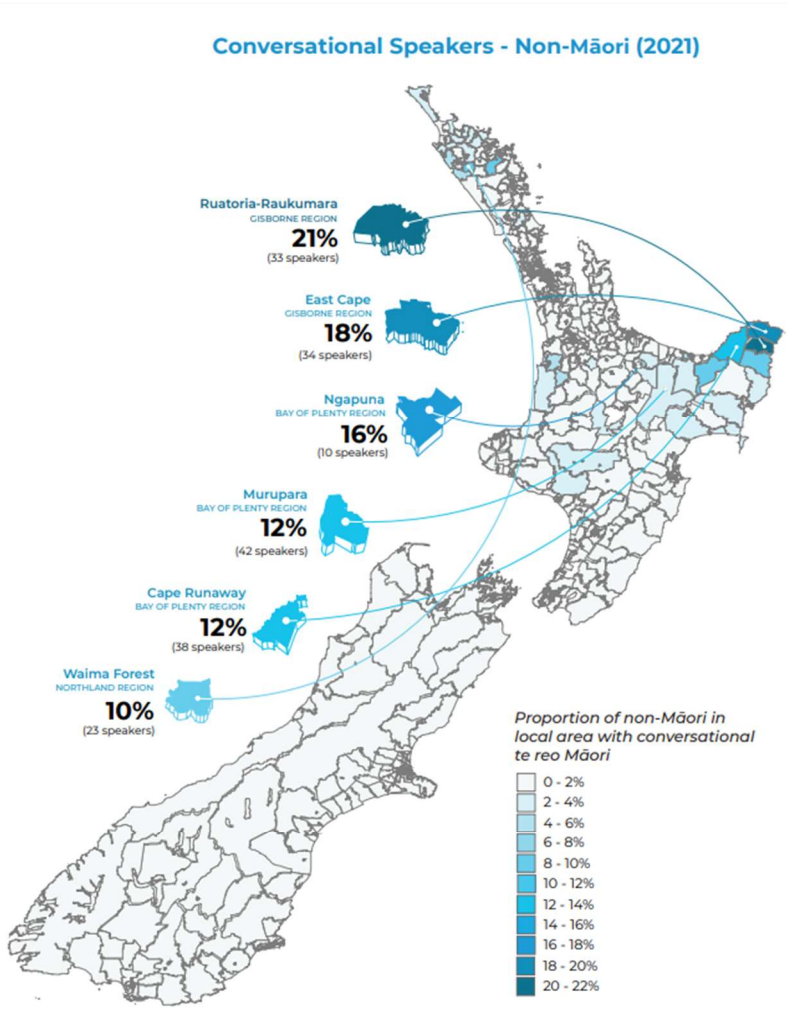


Examples of use of synthetic data at Stats NZ

He Ara Poutama mō te reo Māori allows us to understand more about who the speakers of te reo Māori are – how old they are, if they are Māori and where they live. By measuring the density of speakers at a community level, which is around 1,000 people, we can show that te reo Māori is local.

For example, 26% of the Māori population of the Gisborne region can hold a conversation about everyday things in te reo Māori, however, when we look at the communities in the region we can see that there are geographical areas where there are more speakers.





Plans for future use of synthetic data at Stats NZ

- Investigating possible adjustments to our confidentiality methods -- to be, either more partially differentially private, or fully differentially private

- Investigating possible use of multiple imputation, as part of our editing and imputation methods, as another method of data synthesis

- Investigating the possible use of generative AI based future data platforms to create synthetic data, e.g., databricks

Questions?