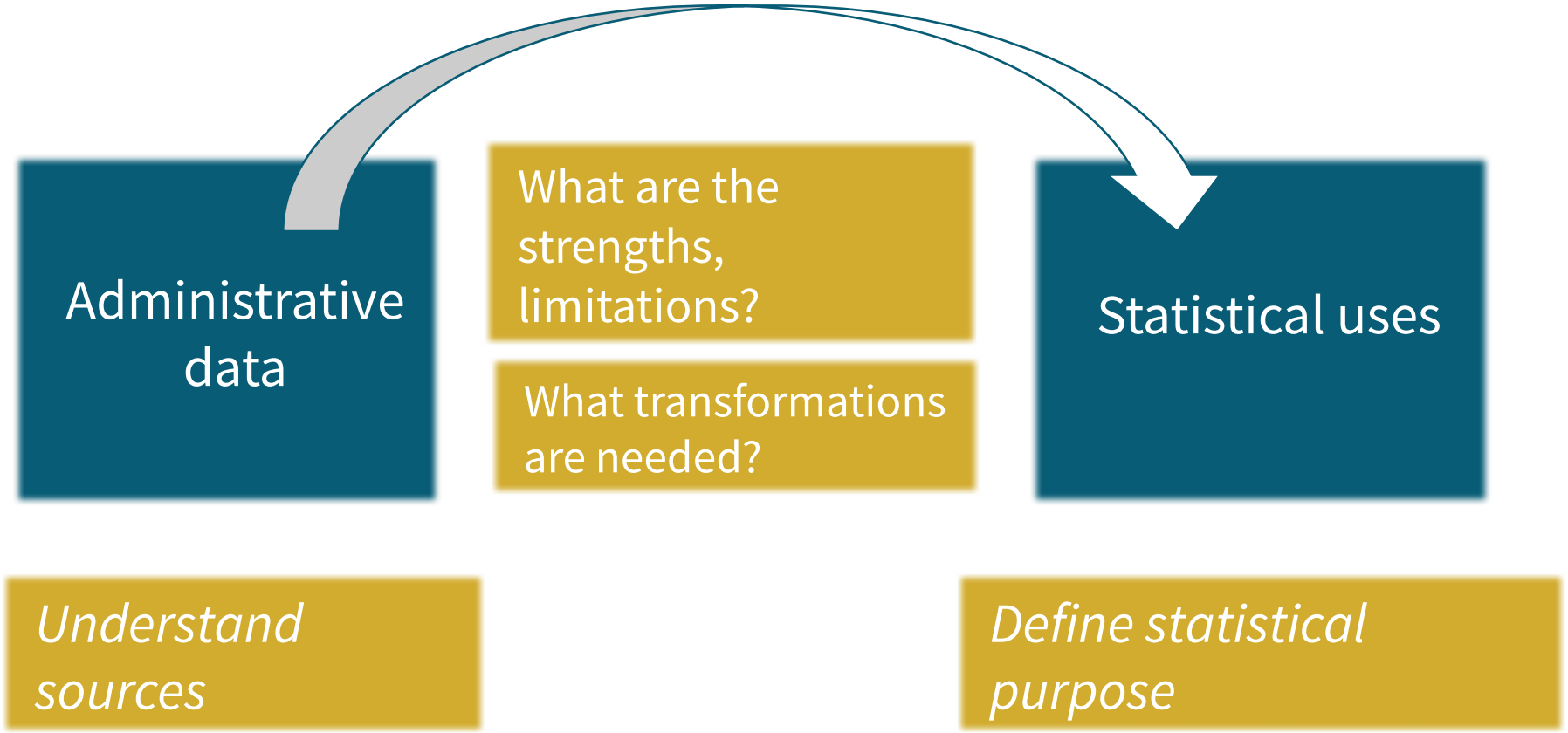
Deriving census attribute information from admin data

NZPopCon, Auckland - 29 August 2023

Hannes Diener & CT team, Stats NZ

Using admin data means a transformation



Admin data quality frameworks guide assessment of variables

Quality Dimensions

- Relevance
- Accuracy
- Timeliness
- Accessibility
- Coherence and Consistency
- Interpretability

Accuracy Dimensions

Representation (units)

- coverage
- unit error

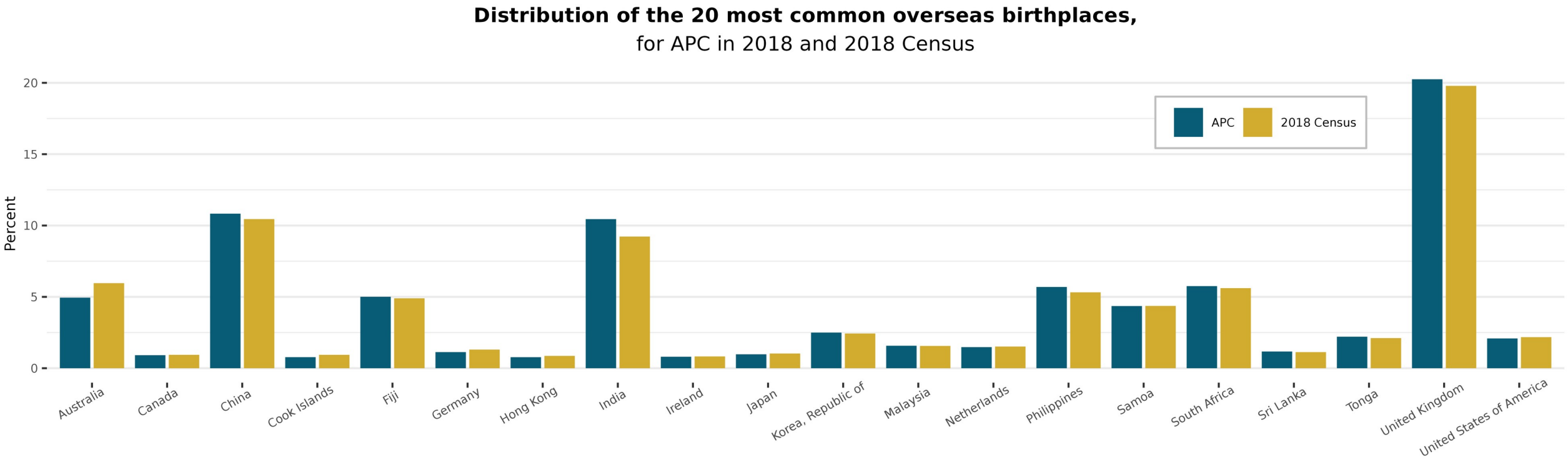
Measurement (attributes)

- validity error
- measurement error (consistency with census responses)

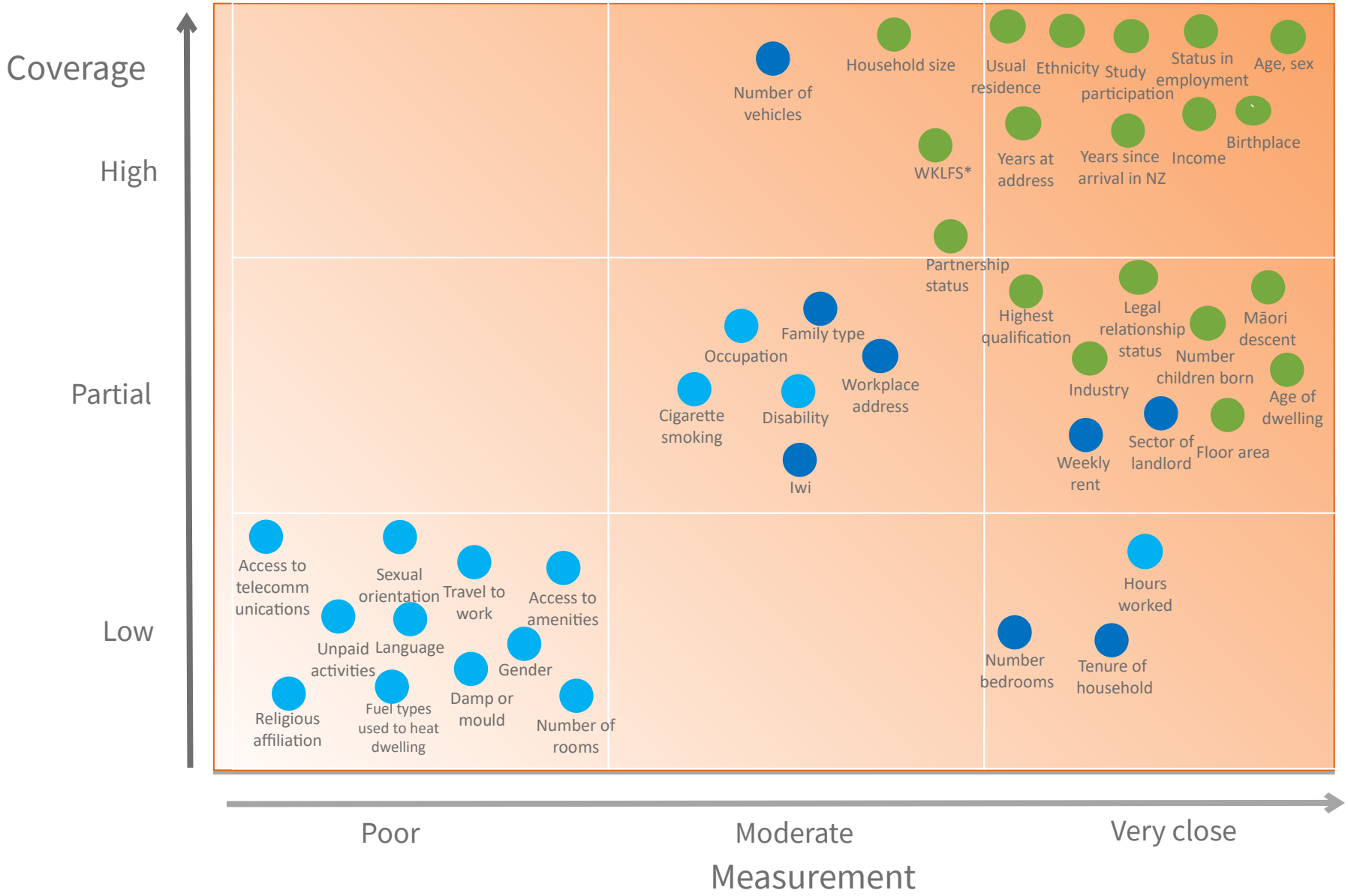
Statistics New Zealand (2016) Guide to reporting on administrative data quality | Stats NZ
Zhang, L-C (2012). “Topics of statistical theory for register-based statistics and data integration.” Statistica Neerlandica 66: 41–63

Example: How country of birth is derived

We use the most reliable source first (this is birth registrations). If there is no data in birth registrations, we use the next best source, if it is available, and so on.



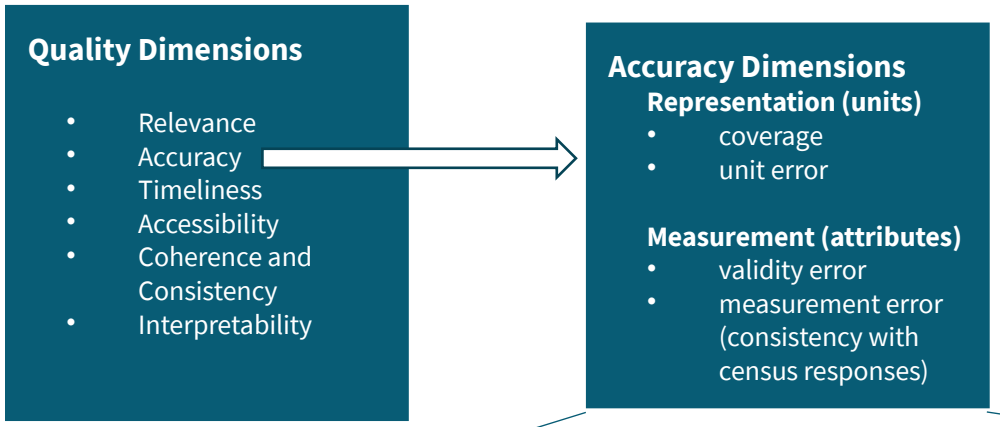
Quality assessment of admin sources



As of Aug 2023. Subject to change

- Metadata
- Detailed analysis
- Experimental series

Quantitative quality measurement



There are a few ways to numerically measure the accuracy dimension:

- Proportion of coverage of attributes **within the admin population**
- **Aggregate comparison** with ERP, Census 2018, and other statistics
- **Unit record comparison** with Census 2018
- Dempster Shafer scores (measuring level of **agreement between several** admin sources)

Research ongoing:

- Use of social surveys for benchmarking in future?
- Model based estimates similar to DS

Quality in the APC

Variable	Missing (percent)	Non-missing quality rating	Output quality
Ethnic group level 1: European	< 1	0.96	0.96
Ethnic group level 1: Māori	< 1	0.94	0.94
Ethnic group level 1: Pacific Peoples	< 1	0.92	0.92
Ethnic group level 1: Asian	< 1	0.96	0.96
Ethnic group level 1: MELAA	< 1	0.79	0.79
Ethnic group level 1: Other	< 1	0.26	0.26
Māori descent	14	0.95	0.82
Birthplace	5	0.96	0.91
Years since arrival in NZ	4	0.93	0.89

Census concepts not well covered in admin data

- Activity limitations (disability) information that is not captured by the health system
- Gender and sexual orientation
- Languages spoken
- Religious affiliation
- Some work and labour force information (‘unemployed/nilf’, occupation, unpaid work)

- Iwi affiliation

Possible *improved collection across the government data system*

Current limitations and mitigations

Error structures are different from traditional census non-response patterns!

- Multiple sources for the same information: we have methods for choosing best values and resolving conflict, especially usual residence address, ethnicity; now extending to models.
- Historical data (pre-digitisation): we already use 2013 Census.
- Sub-groups: for example, migrants/events occur overseas: look for new sources (for example, visas for overseas qualifications).
- Missing categories:
  - Admin has positive identification only: derive ‘No qualifications’, ‘No children’ as default value at 15 yrs.
  - Tax data in Stats NZ excludes ‘No income’: ask for IR zero income data.
  - No sources for unpaid work on family farm or business: do not measure?
- Admin collection issues: for example, ethnicity coding: work with agencies to improve.
- Remaining missing data: develop statistical imputation methods.

He pātai?

Ngā mihi nui

Opportunities, that we can see

- Frequency + Detail: Outputs every year (or more often), with all variables available for all reference dates.
  Already a 16 year time series 2006 -2021

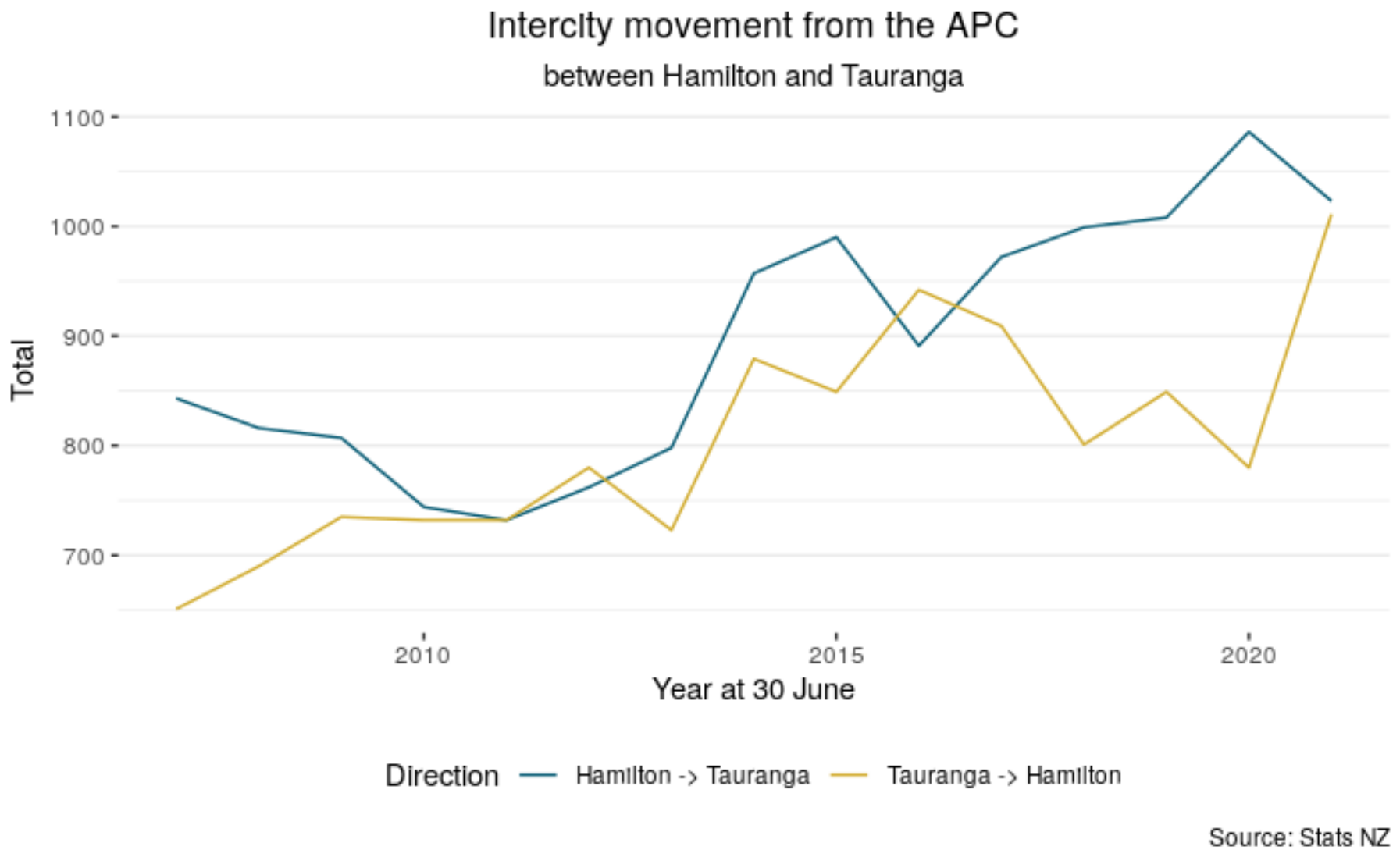
- Quality: More precise, more detailed information, or more accurate – not constrained by respondent's ability to answer questions.

- Resilience: Inherently less exposed to extreme risks or system failure.
  For example, easier to recover from earthquakes, pandemics.

- Longitudinal: The unit record data is longitudinal – this is immensely powerful and largely untapped.
  For example, detailed analysis of flows, cohorts, generational effects.
  Measurement of over-time concepts rather than only point-in-time ones.

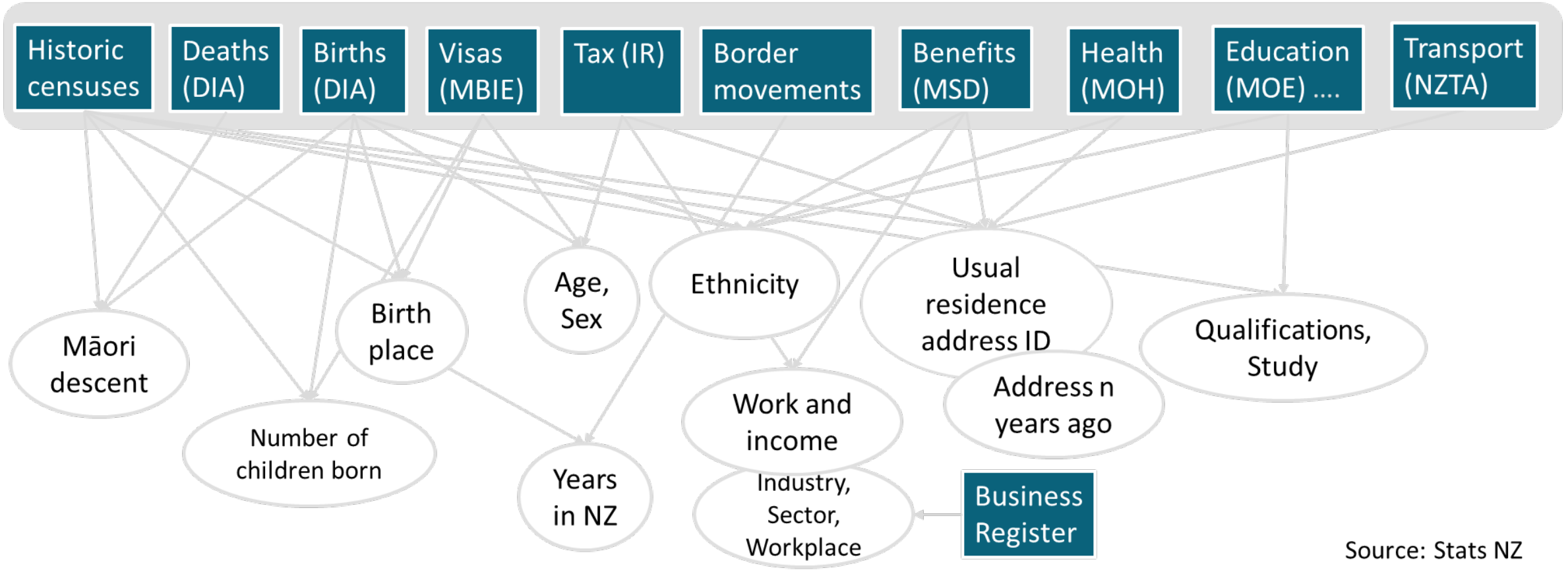
- Synergies with other research: Potential to be a multiplier for other research.
  - Providing common baselines for researchers in the IDI.

Opportunities: longitudinal data



Where does the data come from?

All the data sources are linked in the IDI

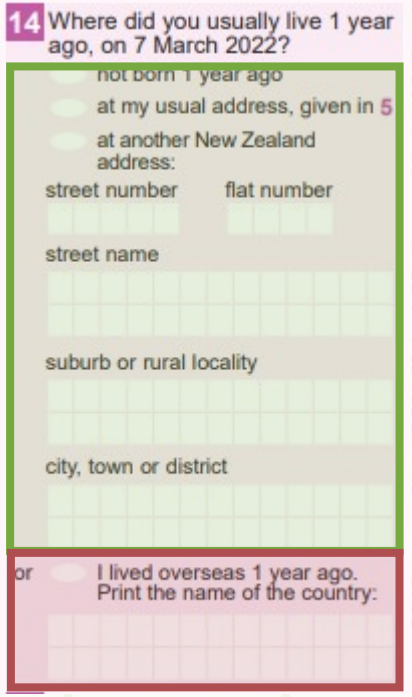


Source: Stats NZ

Individual form census questions in APC

Very **rough** approximation, which does not take into account:

- importance of questions,
- questions that need to be asked on any survey for linking and operational purposes,
- range of the orange category, and
- some details within questions.

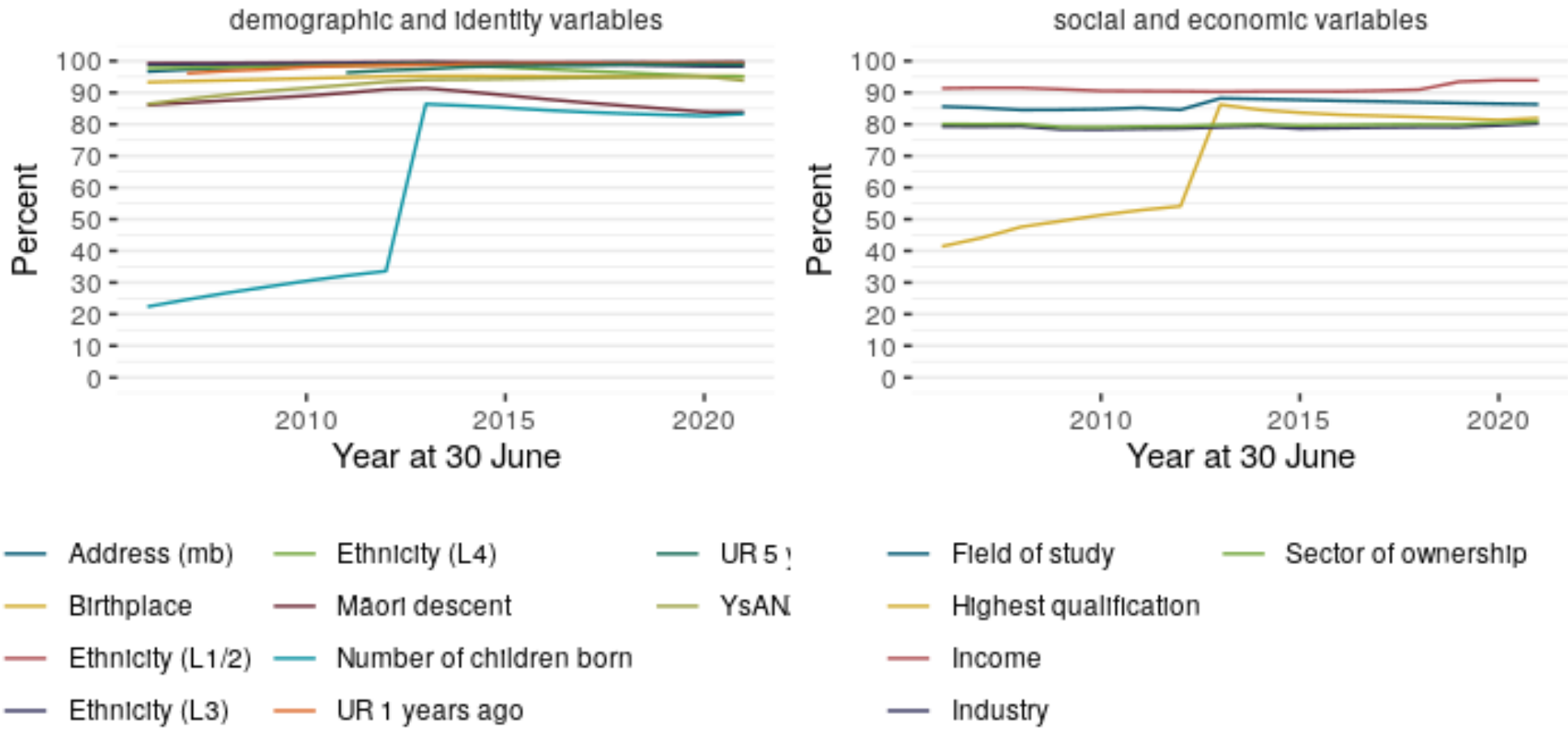


concept part of APC

concept potentially available from admin data

concept unlikely to be available from admin data

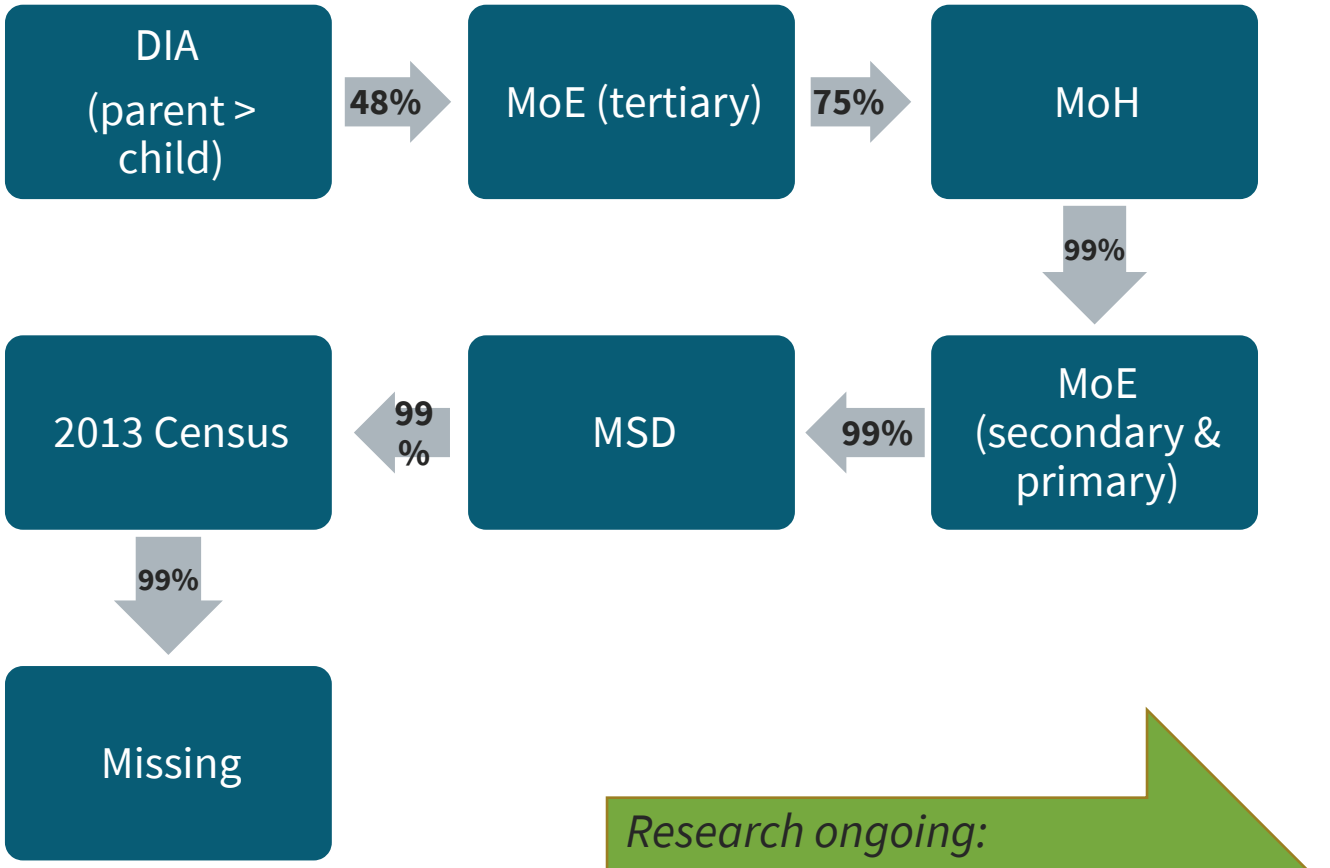
Coverage for each APC variable, 2006 to 2021



Source: Stats NZ

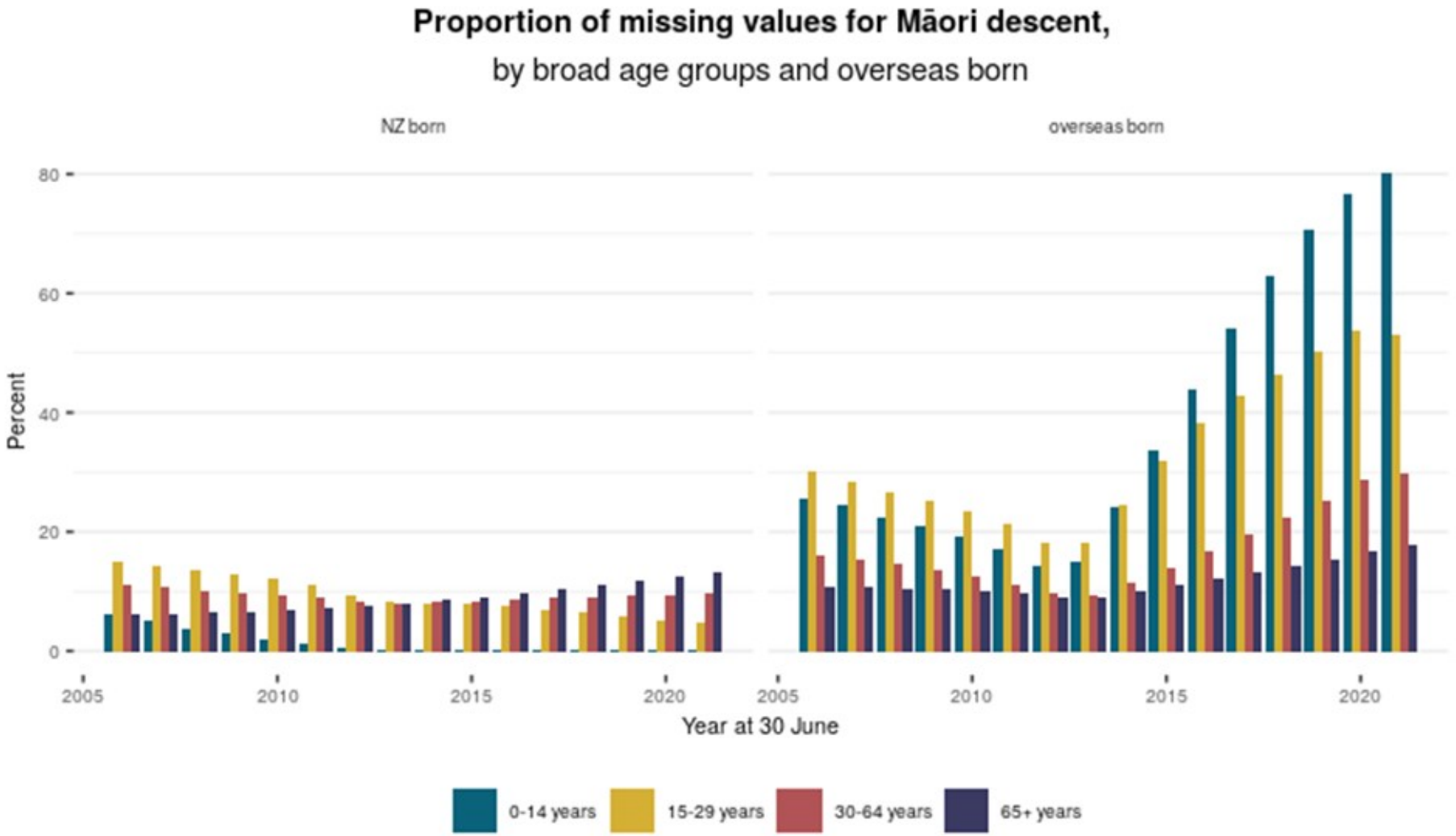
How ethnicity is derived

We use the source that is most similar to the census first (this is birth registrations).
If there is no ethnicity from birth registrations, we use the next best source, if it is available, and so on.
We get an (L2) ethnicity for nearly everyone (99 percent) from the top four sources.



Research ongoing:
- Refining method
- Latent class modelling

Quality: Māori descent



Dempster Shafer theory

Compare selected value, with available values in sources (not necessarily the same sources as used.)

- The more sources, the higher the score
- The more agreement, the higher the score
- The more sources disagreeing, the lower the score

These values are added in the unit record data for:

- Māori descent
- Birthplace
- Years since arrival in NZ
- Level 1 ethnicities (all 6)

Scores for any breakdown can be produced simply by averaging the individual scores!

Dempster Shafer theory

Snz_uid	APC	Source 1	Source 2	Source 3	APC QM↓
12345	A	A	A	A	.95
23456	B	-	B	-	.9
34567	C	C	C	A	.7
45678	A	B	-	A	.4
56789	B	A	B	C	.3
…					