Australian Bureau of Statistics

PROGRESS ON CTC DATA LINKAGE IMPROVEMENTS

Direct measure of income refinement working group paper

January 2021

FOCUS OF DATA LINKAGE IMPROVEMENT WORK

1. As part of the suite of Direct Measure of Income (DMI) refinement work program, the ABS is undertaking the following work with an aim to further understand the reasons why unlinked records did not match to the MADIP spine and improve linkage quality in future Capacity to Contribute (CtC) cycles:

1. investigate linkage outcomes and the characteristics of Address Collection records which did not link in order to inform potential solutions to further improve linkage rates;
2. implement an automated non-standard geocoder for addresses in Aboriginal and Torres Strait Islander Community localities;
3. review and update the names index used to standardise and match given names and surnames, to account for recent new names and cultural diversity changes in Australia.

2. The proposed work has been scheduled for the 2020-21 financial year and it is anticipated that improvements will be implemented for the annual linkage process for the 2021 CtC cycle, where possible.

3. This paper provides an update on progress being made to the linkage improvement work outlined to the DMI Refinement Working Group at the November meeting.

1. INVESTIGATE LINKAGE OUTCOMES TO INFORM POTENTIAL SOLUTIONS

4. The linkage of the 2020 Address Collection to the Multi Agency Data Integration Project (MADIP) Person Spine achieved a linkage rate of 90.8% (for links of an acceptable quality). Given the data available for linking (name, address) this is a very good linkage rate. Further improvements to the linkage rate require reviewing the Address Collection records that did not link to the MADIP Person Spine, assessing the linkage rates by geographic areas and schools with lower linkage rates, and considering the potential for additional data sources to be used in the linking process.

Address collection records that did not link to the MADIP Person Spine

5. To produce high quality linked analytical datasets a matched pair must be unique to be accepted as a link. That is, to make a successful link a single CtC record needs to match with only one Spine record and vice versa. Of the 152,449 (9.2 percent) 2020 Address Collection records that did not link, or link with an acceptable quality, to the MADIP Person Spine, 55,298 (3.3%) had unique matches to the Spine, but were deemed low quality links, 76,793 (4.6%) formed non-unique matches across datasets, and 20,358 (1.2%) could not establish a match using the available linkage variables (Table 1).

6. Of the 76,793 records that formed non-unique matches across datasets, 80% had potential links with more than one MADIP spine record, meaning the linkage process was unable to determine which unique MADIP spine record should link to the corresponding Address Collection record based on the available linkage data. The remaining 20% of non-unique matched records identified the reverse, whereby a singular MADIP spine record showed potential matches to more than one

Address Collection record, and as a one-to-one match had not been achieved, the records were not accepted as a successful link.

Table 1: Breakdown of 2020 Address Collection population when linked to MADIP spine

Number of Address Collection records	Percentage of Address Collection records	Description
1,503,925	90.8%	Records linked to MADIP spine with acceptable quality (quality 1 and 2) – accepted linkage rate
55,298	3.3%	Records that had unique links to MADIP spine that were deemed low quality (quality 3) – excluded from accepted linkage rate
76,793	4.6%	Records with potential links that formed non-unique matches between CtC records and MADIP spine records
20,358	1.2%	Records where no link could be established
1,656,374	**100%**	**Total**

7. Table 2 outlines an example of how names within the same geographic area can prevent unique matches from being formed. Examples can be seen within the same household, such as parents and children having the same first name and surname. Examples can also be seen at higher geographic levels such as Statistical Area Level 1, which come into the linkage process where Address Collection records do not match to MADIP spine records on lower levels of geography.

Table 2: Non-unique match example: Common Name Agreements

CTC ID	Spine ID	Geography	First Name	Surname	Date of birth (Spine)
B	1	9999	John	Smith	1/1/2000
B	2	9999	John	Smith	3/4/1990
B	3	9999	John	Smith	5/5/1980
B	4	9999	John	Smith	12/12/1970
B	5	9999	John	Smith	2/8/1995

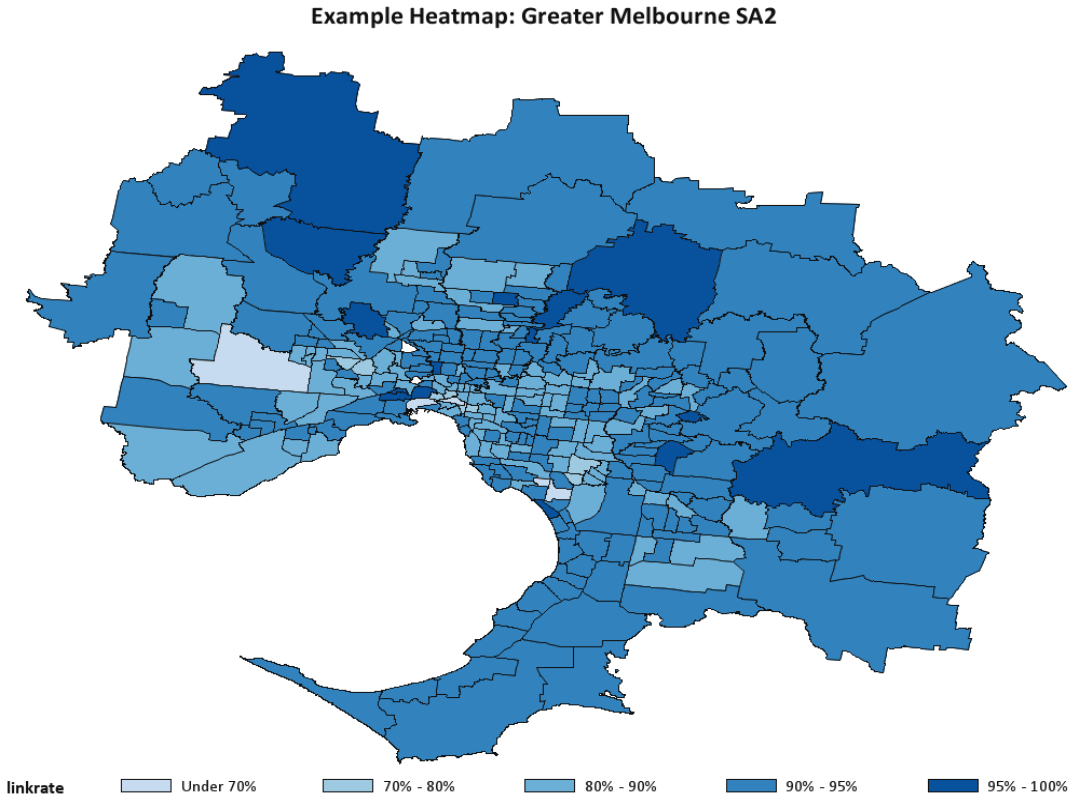
8. The next steps for this analysis will involve understanding the prevalence of non-unique matches at the household level versus higher levels of geography in order to inform possible solutions for reducing non-uniqueness during the linkage process, and in turn, further enhance linkage quality.

Analysis of linkage rates by geography

9. The ABS have commenced analysis of linkage rates by geographic level to assess for noticeable areas where linkage rates may vary, using heatmaps of geographic areas of parental addresses (see below for the Greater Melbourne Statistical Area Level 2 geographical regions by 2020 CtC linkage rate bands).

10. This aims to assist understanding of patterns in linkage outcomes, and inform the focus of subsequent investigations. The most obvious pattern is the urban/rural divide in linkage rate levels. Linkage rates are generally highest in the inner cities of the State capitals. Early analysis indicates

that, aside from the urban/rural split, overall linkage rate variations appear to be random on the heatmaps with no obvious geographic clusters of unlinked records.



11. The next steps for this analysis include identifying lower linkage rates by geography, particularly remoteness indicators, that have been consistent across the 2018, 2019 and 2020 linkage cycles of the CtC program. Analysis of school-level linkage rates by geographic region is also planned for this body of work.

Australian Electoral Commission data as possible linkage enhancement source

12. Without other linkage variables in the linkage process, the CtC linkage is reliant on high quality and up to date address information in both the Address Collection and MADIP. The ABS has identified data from the Australian Electoral Commission (AEC) as a possible data source to augment the linkage process for future iterations of the CtC linkage. It is likely that address information for some parents or guardians may be more up to date on the Electoral Roll than the core MADIP Person Linkage Spine datasets, particularly in regions where there has been a recent election.

13. Given address is a critical linkage item for the CtC program, this AEC dataset has the potential to improve linkage outcomes for Address Collection records that did not successfully link on low levels of geography to a corresponding MADIP spine record. The ABS is currently assessing the extent to which an AEC dataset will support the CtC project objectives, with a recommendation to be provided at the end of this assessment.

2. NON-STANDARD GEOCODER

14. For CtC 2020, the ABS employed a manual process whereby residential addresses that failed to match to a location on the ABS Address Register were manually mapped to a separate list of Aboriginal and Torres Strait Islander Community localities. Over 2700 parent records were coded from this separate list to enable use in the linking process.

15. In addition to automating this process and implementing it before the 2021 linkage cycle of Address Collection records, the ABS are investigating process improvements that aim to better match Community addresses across CtC and MADIP spine records. Improvements have been made to the program for matching by adding the capability for fuzzy matching with a user specified threshold. This has shown a significant increase in matching rate when matching Community names against suburb in CtC.

16. The next steps are to test and report on any possible improvements identified as part of this work. The productionised program will be completed and tested prior to live implementation, before linkage of the 2021 Address Collection.

3. REVIEW AND UPDATE NAMES INDEX

17. An individual's name is a key variable for data linkage in CtC, as name and address are the only available linkage variables. It is therefore important to have both high-quality name data from the CtC Address Collection, as well as robust processes to find valid matches between records across the two different data sources. As the cultural diversity of Australian society changes and evolves over time and new names become more common among the Australian population, it is important to be able to incorporate associated changes in naming conventions and patterns in the linkage process.

18. The ABS have conducted a literature review of information sources, such as academic articles, books and media, relevant to the practice of processing and standardising names. The research has identified a number of possible data sources that can be used to update the names index, which aims to create a more comprehensive index that accurately accounts for the diversity in Australia's society. It has also but also verified that our current index-based method is considered best practice right now.

19. The next steps include producing a prototype updated index, and comparing linkage results using the old and new indexes. The engagement of cultural consultants may be considered pending testing results of the updated names index and analysis of unmatched name records.