Integrated Research Infrastructure for Social Science Project Report

ARDC HASS RDC and Indigenous Research Capability Program

2.2 IRISS Milestone_WP2. Technical report 1-Longitudinal inconsistencies in categorical Census data

Authors: Matthias Kubler

2022-09-30

LEAD ORGANISATION: Australian National University (ANU), Australian Data Archive (ADA), ANU Centre for Social Research and Methods (CSRM), Research School of Social Sciences (RSSS)

PARTNER ORGANISATIONS:

- University of Queensland, Institute for Social Science Research (ISSR)
- University of Melbourne, Melbourne Institute
- Australian Urban Research Infrastructure Network (AURIN)
- Australian Consortium for Social and Political Research Inc (ACSPRI)

Longitudinal inconsistencies in non-spatial categorical classifications in Census data

Introduction

This document outlines some of the typical changes affecting categorical variables in the Census data collections that can occur over time. Types of changes are illustrated using examples, which relate to changes between the 2016 and 2011 Censuses. This is accompanied by brief discussions of the documentation of such changes in ABS materials and how changes could be addressed when analysing data across Censuses.

The ABS documents changes to Census variables, whether triggered by changes in the data capture or the data processing, in a 'What's New for <year>' section, which is part of the respective Census Dictionary for that year. The examples given to illustrate types of changes in this document were sourced from such a section in the 2016 Census Dictionary (https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/2901.0Main%20Features202016?opendocument&tabname=Summary&prodno=2901.0&issue=2016&num=&view=)

The 'What's New for <year>?' sections do not fully document changes, which will be pointed out in this document. To better illustrate changes to variables, screen shots from relevant sections of the 2011 and 2016 Census Dictionaries are included in the presentation below. Where quotations are used in the document these are from the ABS and relate to the 'What's New for 2016?' section.

Types of change

This section outlines different types of changes. This starts by presenting types of changes that are more difficult to detect or to assess.

Change in mode of participation/collection

Census data collections have been moving towards online administration over the past three Censuses. In 2011, about one third of Census completions (at the household level) were undertaken online, in 2016 about two thirds. This was expected to rise to 75% in the 2021 Census. Offering dual mode completion has been associated with the ABS implementing changes in wording and layout between the paper and online versions of the household questionnaire to optimise the questionnaire for the online environment, but also generally: "The development of the online questionnaire for 2016 has provided an opportunity to make refinements to gain more accurate data from respondents, while decreasing the burden placed on those filling out the form."

Identifying differences between online and paper versions of the Census questionnaires may require independent investigation. Assessing how such changes affect responses will be hard to quantify.

Changes in the mode of participation may have also affected how some information is captured/collected: "The move to a new method of conducting the Census also meant a change to how data on Dwelling Location (DLOD), Dwelling Type (DWTP), Structure of Dwelling (STRD) and Type of Non-Private Dwelling (NPDD), previously recorded by Census collectors, are obtained." While the ABS goes on to provide more information on the change for the variables mentioned, this information does not easily make apparent what the change consisted of, without a more intimate understanding of Census data collections over time:

"There has been a change in the way this information is collected for 2016. It was recorded by ABS Address Canvassing Officers in the lead up to the Census as part of establishing the Address Register as a mail-out frame for designated areas. In areas enumerated using the traditional approach of delivering forms, the information was collected by ABS Field Officers during the Census collection period. Dwelling type was also updated as required by ABS Field Officers during the 2016 Census enumeration period."

Is a Census Collector equivalent to an ABS Field Officer? And if Address Canvasing Officers recorded the information before the Census in 2016 as opposed to Census Collectors during the Census in 2011, did both use the same observational code frame for recording the information?

Users of ABS Census data, particularly users of time series or longitudinal Census data should be alerted to such differences between online and paper version or changes in the way that ABS staff collect information that is included in the Census.

Change in question wording while retaining response options

Independent of moving the Census data collection to an online environment, the ABS sometimes makes (slight) changes to question wording between Censuses that can be hard to pick up when variable names, value categories and value labels remain the same. At times, such changes consist of changes to secondary guidelines, such as giving examples of acceptable entries in open-ended fields.

There were several changes to question wordings or accompanying instructions in the 2016 Census. The ABS document such changes, often in a descriptive format as is shown in the example below.

Example: Variable Highest Year of Schooling completed (HSCP)

Census 2016
"A minor change was made to the dot point instruction in the Census question, to clarify that people attending school should mark the last year completed not the current year of study."

Users can further clarify which change took place by visiting the Census Household forms from 2011 and 2016. However, presently they need to do this by their own initiative unprompted by the ABS documentation. It would help if this possibility was made explicit in the documentation, or, even better, if the documentation of the change included the 2011 and 2016 questions (e.g. screenshots of relevant parts of the Household form).

The documentation would be further enhanced if it contained some observations or even speculations of the impact the change could have had on the data.

Change in variable label

A change with usually minor implication for data integration is a change in a variable's label. The example shows the variable GNGP, which was called Public/Private Employer Indicator in the 2011 Census and was relabelled to Public/Private Sector in the 2016 Census.

Example: Variable GNGP

Census 2011	**Census 2016**
Public/Private Employer Indicator	Public/Private Sector

The resulting discrepancy could be addressed by aligning the variable label in the data for both years (if that was beneficial in some data analysis process).

Change in value label (only)

Labels for individual categories of a variable can also change between Censuses.

In the example below, the value label for category 1 in the Indigenous Household Indicator changed in 2016. There is no further information about this change, so that one can suspect that there was no other change associated with that change in the category's label, such as a change in the underlying question(s) on the Census form and/or a change in the data processing rules when deriving the category. In the case here, changing the 'Indigenous' label to 'Aboriginal and/or Torres Strait Islander' makes the content of the category more specific and reflects shifts in using such terminology in other data collections, so that it

appears plausible that this was the only change. However, ideally, the user should not be left with even a slight sense of ambiguity about what the change may have entailed as it is common that changes in a label of a category indicate a change in the content of the category.

Assuming the change in wording of the category label was the only change, the category means the same in 2016 than in 2011 (assuming no impact from changes in participation mode in 2016), and the discrepancy in the data could be addressed by aligning the value label across time (if that was beneficial in data analysis processes).

Example: Variable Indigenous Household Indicator (INGDWTD)

Census 2011	Census 2016
Value label for category 1: Indigenous	Value label for category 1: Aboriginal and/or Torres Strait Islander

Change in category - splitting of a category

In this scenario a previous category is split into multiple categories. Below is a simple example for the variable Dwelling Structure, which had one category that combined Caravan, cabin and houseboat in 2011, and two categories that covered these three options in the 2016 Census.

Example: Variable Dwelling Structure (STRD)

Census 2011	Census 2016
Value 91 - Caravan, cabin, houseboat	Value 91 - Caravan
	Value 92 - Cabin, houseboat

Re-aligning the 2011 and 2016 categories by aggregating the 91 and 92 categories in 2016 to one category could be a solution when temporally consistent categories are required in the analysis.

To add complexity here the variable Dwelling Structure was also affected by a change in how this information was captured as reported further above (under Change in mode), and this change remains somewhat opaque.

Change in content of a derived category

The information included in a category of a variable can change as a result of changes to question wording, response options and/or changes to rules by which variables and their categories are derived from source variables.

A potential example of the latter is given below for the 'not applicable' category of the variable Number of Employees. 'Potential' is used here as it is not entirely clear whether the variable is derived from multiple variables. Question 37 asks people who work in their own business 'Does the person's business employ people?' providing three options:

- No, no employees (other than owner/s)
- Yes, 1-19 employees
- Yes, 20 or more employees

The reported categories include the three categories above, which are directly taken from the responses to the question. However, as the question is asked of a sub-population it is likely that the ABS checks responses to the preceding questions to derive those that should be coded as 'not applicable' independently, without fully relying on responses to Q37. This could entail changing a response given for one of the three categories to 'not applicable' after determining that a respondent who gave a response should not have given one.

In 2016, the ‘not applicable’ category included persons who had not stated their employment status. Again, the ABS documentation in the data dictionary leaves open whether this condition was just added to the dictionary as it had been forgotten previously or whether the addition also signified adding a condition for coding to the ‘not applicable’ category that was not in place in 2011.

The ABS’s documentation of the change “'Not applicable' has the additional category of 'Persons with Status in Employment (SIEMP) not stated'.” does not remove this ambiguity.

If the change entailed a change in derivation rules, the user should be informed about the derivation of the variable in both years to such a degree that they can independently derive the Number of Employees variable in the 2011 and 2016 data, and investigate what difference the change in 2016 would have made in 2011 or vice versa, what difference to the 2016 data applying the 2011 coding rules would have made. In this scenario, the data integration solution could consist of the user newly deriving the variable for 2011 or for 2016 so that it is consistently derived in both years.

Example: Variable Number of Employees (EMPP)

Census 2011 Not applicable category	**Census 2016** Not applicable category
• Employees • Contributing family workers • Unemployed persons • Persons not in the labour force • Persons with Labour Force Status (LFSP) not stated • Persons aged under 15 years	• Employees • Contributing family workers • Unemployed persons • Persons not in the labour force • Persons with Labour Force Status (LFSP) not stated • Persons with Status in Employment (SIEMP) not stated • Persons aged under 15 years

Treatment of the highlighted status in deriving the ‘not applicable’ category in 2011 is not clear.

Note that there was another change in 2016 that could have affected the resulting variable in 2016: the question instructions changed so that owner managers were instructed to exclude themselves from the count of people that they employ. This is a change that would fall under ‘Change in question wording’ discussed further above. It would be hard to assess the impact of this change using only Census data as the number of employees is only captured in ranges. Some external reference data source that covers the 2011-16 period, such as administrative business registers could help in such endeavour.

Change in categories’ (dollar) ranges

It is not uncommon that dollar ranges are updated for relevant variables (e.g. affecting personal and/or household income or rent/mortgage payment variables) in the Census. The example shown here is for Total Personal Income (weekly). The highlighted categories in the Census 2011 column do not exist in the Census 2016 column and vice versa, the highlighted categories in the Census 2016 column do not exist in the 2011 Census data.

The data integration solution in this case could be to aggregate the 2011 and 2016 Census categories so that they are consistent, which is possible in this example by:

- combining the 2011 categories 03 and 04 to create a category for the range $1-$299, which can be replicated in the 2016 data by aggregating the 2016 categories 03 and 04; and
- combining the 2011 06 and 07 categories to create a category for the range $400-$799, which can be replicated in the 2016 data by aggregating the 2016 06, 07 and 08 categories; and
- combining the 2011 categories 11 and 12 to a category with the range $1500 or more, which can be replicated in the 2016 data by aggregating the 2016 categories 12, 13, 14 and 15.

While this would achieve consistency, it would also reduce the level of detail and variation in values when undertaking data analyses. One question for a user of the data is whether the change in categories’ ranges was created post-data collection or when the data was captured. In the former case, the user could still

mount a data request to get the underlying data and create their own alternative consistent ranges. Again, the ABS documentation of the change "The categories for personal income dollar ranges have been revised for the 2016 Census." is not detailed enough to alert the user to how the change was undertaken. Currently, users need to consult the respective data dictionaries and Census forms to independently find out more about the changes between Censuses.

Example: Variable Total Personal Income (weekly) (INCP)

Census 2011	Census 2016
01 Negative income 02 Nil income 03 $1-$199 ($1-$10,399) 04 $200-$299 ($10,400-$15,599) 05 $300-$399 ($15,600-$20,799) 06 $400-$599 ($20,800-$31,199) 07 $600-$799 ($31,200-$41,599) 08 $800-$999 ($41,600-$51,999) 09 $1,000-$1,249 ($52,000-$64,999) 10 $1,250-$1,499 ($65,000-$77,999) 11 $1,500-$1,999 ($78,000-$103,999) 12 $2,000 or more ($104,000 or more) && Not stated @@ Not applicable VV Overseas visitor	01 Negative income 02 Nil income 03 $1-$149 ($1-$7,799) 04 $150-$299 ($7,800-$15,599) 05 $300-$399 ($15,600-$20,799) 06 $400-$499 ($20,800-$25,999) 07 $500-$649 ($26,000-$33,799) 08 $650-$799 ($33,800-$41,599) 09 $800-$999 ($41,600-$51,999) 10 $1,000-$1,249 ($52,000-$64,999) 11 $1,250-$1,499 ($65,000-$77,999) 12 $1,500-$1,749 ($78,000-$90,999) 13 $1,750-$1,999 ($91,000-$103,999) 14 $2,000-$2,999 ($104,000-$155,999) 15 $3,000 or more ($156,000 or more) && Not stated @@ Not applicable VV Overseas visitor

Change in category order

The example here relates to the variable Level of Highest Educational Attainment. The variable is derived from questions on non-school and post-school education and the derivation rules define the order of the categories. In 2016 the order of the categories was changed to align with ASCED. The change consisted of moving the Certificate Level I and II to between Secondary Education Year 9 and Year 10. This was associated with breaking down the previous higher-level category of 'School Education Level' into 'Secondary Education – Years 9 and below' and 'Secondary Education – Years 10 and above'. To reflect the new sequence of educational levels, the numerical value codes of the categories as well as of some of the higher-level categories, were changed in this process. The example shows an extract of the categories that were affected by the change.

As in other cases, the ABS documentation in the 2016 Census data dictionary is not overly specific in talking about the change: "Categories within the HEAP variable have been re-ordered to align with the Education standard. In particular, non-school qualifications Certificate III and above are listed above Year 12 and Certificates I and II are listed below Year 10." For someone unfamiliar with the questioning on the Census form it leaves open whether the re-ordering was achieved by changes to the question(s) or changes in data processing.

From visiting both, the 2011 and 2016 household forms, we know that the questions remained the same (in the case of non-school qualifications open-ended questions that were coded to ASED), so the re-ordering was achieved in the data processing. The data integration solution in this case could be to align the sequencing of categories by changing some of the numerical codes.

Note, however, that in this example, there is a 2011 category 500 – Certificate Level, nfd that has, according to the data dictionary, no equivalent in 2016. Is it possible that responses coded to this category in 2011 would have been coded to 001 Inadequately described in 2016? Regardless, the ABS do not appear to clearly document what happened to category 500.

Example: Variable Level of Highest Educational Attainment (HEAP)

Census 2011	Census 2016
5 Certificate Level	5 Certificate III & IV Level
50 Certificate Level, nfd	510 Certificate III & IV Level, nfd
500 Certificate Level, nfd	511 Certificate IV
51 Certificate III & IV Level	514 Certificate III
510 Certificate III & IV Level, nfd	6 Secondary Education - Years 10 and above
511 Certificate IV	611 Year 12
514 Certificate III	613 Year 11
52 Certificate I & II Level	621 Year 10
520 Certificate I & II Level, nfd	7 Certificate I & II Level
521 Certificate II	720 Certificate I & II Level, nfd
524 Certificate I	721 Certificate II
6 School Education Level	724 Certificate I
611 Year 12	8 Secondary Education - Years 9 and below
613 Year 11	811 Year 9
621 Year 10	812 Year 8 or below
622 Year 9	
067 Year 8 or below	

New derived variable

Another type of change that can occur in Census data and reporting is the introduction of new variables that are derived from source variables.

The example below relates to a variable that was introduced in the 2016 Census data. The variable introduced in 2016 expresses different levels of engagement in education and/or the labour market.

Example: Variable Engagement in Employment, Education and Training (EETP)

Census 2011	Census 2016
Non-existent	Derived from data items Labour Force Status (LFSP), Hours Worked (HRSP), Full-Time/Part-Time Student Status (STUP) and Age (AGEP)

This variable could be created the same way in the 2011 Census data, if that was beneficial for data analysis. While the Glossary of the Census Dictionary 2016 includes a description of each category it does not include the specific coding rules and includes a reference to the National Information and Referral Service: "For the 2006 and 2011 Censuses, data for this item can be derived based on existing data items - contact the National Information and Referral Service (NIRS) for this data." The NIRS is a consultancy service. Referencing it here suggests that the ABS does not anticipate that users of their data products would or should independently create the variables in previous Census data (e.g. after extracting data using TableBuilder). Such assumption would be consistent with the descriptive rather than specific/prescriptive character of the ABS documentation of the variable's categories.

Summary

This document outlined some types of changes that affect the consistency of available Census data that have occurred between Censuses using some changes introduced in the 2016 Census as illustrative examples. These included changes to questions, variable and category labels, changes to category content via splitting of a previous category or changes to derivation rules, and changes to the order of categories

(and their numerical codes). There will be various other types of changes that have not been considered in this brief examination.

Notwithstanding the incompleteness of covering all types of changes, there are some general issues/points that arise from the exercise.

ABS Census documentation (2011-2016[1])

a) The ABS makes available a number of resources data users can peruse to identify and better understand changes it introduced, most notably:
   - Census data dictionaries, which include a 'What's new…?' section, sections for individual variables and a Glossary with further information on variables or broader concepts that relate to multiple variables (e.g. income);
   - Census household forms that show the underlying questions and response options and skipping patterns used to capture information; and
   - References to documentation of larger classifications, such as for countries, religions, languages, educational qualifications, industry and occupation.

b) With the exception of the larger classifications, which are referenced and linked and which contain documentation about changes, the onus is on the user to identify and search these materials for the different Census years independently to further scrutinise changes between two particular Censuses. The need to do so is influenced by the next point.

c) The documentation surrounding changes between Census years in the 'What's new…?' and Glossary sections of the 2016 Census Dictionary tends to be descriptive and insufficient to understand changes in technical detail necessary to contemplate data integration issues and solutions.

d) Some questions that users may have in the context of understanding changes can be pieced together from scrutinising Census dictionaries and Census questionnaires for different years. Others, which require knowledge of detailed coding or derivation rules cannot.

e) There tend to be no statements about how changes to the Census data collection (could) impact on the data. There is perhaps an implied assumption that changes (e.g. to wording of a question or instruction) would not significantly impact.

Overall, there is a lot of documentation of data for individual Census years. The documentation of changes surrounding the 2016 Census data is not user friendly as relevant information that is needed to shed more light on changes needs to be identified and compiled by the user from individual sections of multiple Census Dictionaries and/or the associated Census Household forms, which are not linked to in the 'What's new…?' or Glossary sections of the dictionaries. This particularly applies to users who are not familiar with the Census data collection and its questions. Further, changes to the Census data collection or processing tend to be documented in a descriptive and general manner, which can lack sufficient detail for users to fully understand and independently address inconsistencies across Census data collections.

Data user requirements

The exercise undertaken here can also shed some light on user requirements when dealing with temporal inconsistencies in Census (and other) data.

At a minimum, users should be alerted to a change surrounding the capture or processing of the data they are dealing with and should be referred to documentation about the change. The detail of this documentation may vary dependent on the user's capability and interest. Users who want to undertake

[1] The 2021 Census Dictionary appears to include more detail on how information was captured and on what changes took place.

time series or longitudinal analyses that involve variables affected by change, need detailed information about the change.

It would be desirable that the documentation of the change was available in a user friendly and easily accessible format. It would further be desirable for the documentation to include an expert assessment of the impact the change would, or could, have on the relevant information.

Ideally, users of data affected by temporal inconsistencies would also receive recommendations about how to deal with the inconsistencies under different scenarios, whether that entailed possible ways of independent investigation of the impact of the change that the user could undertake, disclaimers for the interpretation of results, or procedures/strategies for harmonising the data from different years that the user could pursue. This could be accompanied by data integration tools, such as machine-readable concordance tables or scripts in a number of languages.

At the moment, the ABS documentation of changes to categorical non-spatial variables in the 2016 Census does not quite reach the minimum user requirement because the changes are sometimes not documented in sufficient detail and/or the available documentation does not directly refer to relevant other documentation that could clarify some change.