

# Enumeration errors of census imputation

James Chipperfield<sup>1,2</sup> and Li-Chun Zhang<sup>3,4</sup>

<sup>1</sup>Australian Bureau of Statistics, Australia  
(James.Chipperfield@abs.gov.au)

<sup>2</sup>National Institute for Applied Research Australia, University of  
Wollongong, NSW, Australia

<sup>3</sup>Statistisk sentralbyrå, Norway

<sup>4</sup>University of Southampton (L.Zhang@soton.ac.uk)

## Abstract

Nonresponse to census enumeration is unavoidable despite the effort and resource committed. Imputation for the census non-respondents so as to create a complete census database is common practice for preparing the numerous census outputs. Many countries have used administrative records for census imputation or are investigating such uses, in order to benefit from imputing genuine data of the census non-respondents, which would not be possible when the imputed records are taken from the census respondents. But there is a gap in the literature regarding the assessment of the associated enumeration errors. In this paper we develop methods for estimating the enumeration errors induced by census imputation and illustrate their uses in the context of Australian Census 2016.

**Key words:** erroneous enumeration, missing enumeration, linkage error, duplication, coverage error, administrative records

## 1 Introduction

Population counts by demographic and geographic breakdowns are perhaps the most fundamental official statistics. In many countries the trusted counts are calculated from a census (or census enumeration). A major source of error of the census counts is unit nonresponse. Imputation for the records missed by census is preferable to case weighting the census enumerated records for the numerous census outputs, and it has been common practice for statistical agencies to impute for census non-respondents (United Nations 2017a). For instance, in the so-called one-number census (Office for National Statistics, 2001), the census nonrespondents are imputed to produce a complete census database so that all statistics add to ‘One Number’.

When the imputed records are taken from the census respondents, these are assumed to represent the census non-respondents in some predictable way

after controlling for relevant characteristics (e.g. Brown et al, 1999; Farnell and Darby, 2020). Alternatively, the imputed records can be taken directly from the administrative sources available to the statistical agency, the data of which are routinely collected about a significant proportion of residents via their access of government services. An administrative record that is missed by the census will likely belong to a resident if it has associated activity, or ‘sign-of-life’, that is consistent with a resident; whereas such genuine records for census non-respondents cannot exist among the census enumerated records. However, any operational rule or likelihood threshold, such as ‘at least one administrative transaction in the previous 24 months’, can potentially cause both over- and under-counting errors of the imputed census. For a review of this problem see UNECE (2018) and United Nations (2021) and Statistics New Zealand (2019a) for applications.

Notwithstanding the potential errors, Statistics New Zealand (2017, 2019a, 2019b) imputed administrative records for non-respondents in the Census 2018, as the assumptions required of imputing census respondents was judged to be invalid particularly for Māori and Pacific populations. At the Office for National Statistics, Tietz et al (2019) considered imputing the census questions “how many rooms” and “highest qualification” from administrative sources, which are obtained by dwelling and person-level linkage respectively. Many other agencies such as Statistics Canada (2021), Statistics Scotland (2022) and the Australian Bureau of Statistics (2021a) are using or investigating such uses of administrative data.

Nevertheless, there is currently an apparent gap in the literature when it comes to assessing the coverage errors induced by imputing administrative records for census non-respondents. To be specific, let  $U$  be the hypothetical set of records corresponding to the target population. Let  $C$  be the set of census (enumeration) records. Let  $A$  be the set of records originated from the administrative sources. The records in  $C \setminus U$  may be called the *erroneous enumerations* of  $C$  and those in  $U \setminus C$  its *missing enumerations*. Similarly for  $A \setminus U$  and  $U \setminus A$ . While both the sets  $C$  and  $A$  are completely known, their enumeration errors are generally unknown and need to be estimated.

For imputation of the census missing enumerations, one can either use the records in  $C$  or  $A$ . Let  $U_{imp}$  denote the set of imputed records. Any record in  $U_{imp}$  may be said to cause an *enumeration error* if it does not belong to  $U \setminus C$ , which would translate into coverage errors of the imputed census.

The term “enumeration error” is emphasised because the problem differs to the usual “imputation error” given missing data (e.g. Little and Rubin, 2002), where all the units are identified and known but some (or all) of the associated variables may be missing for a given unit. Enumeration error arises due to the ambiguity surrounding the target statistical units (Zhang and Chambers, 2019), rather than missing attributes associated with known units. This is also why we refer to  $U$ ,  $C$  and  $A$  as collections of records rather than persons. Although one aims to enumerate persons by the census, the results of census and the material one can work with are the census (enumeration) records. The census enumeration error, such as erroneous, duplicated or missing enumeration, refers then to the discrepancies between these census records and the target population units, for the latter of which one must as well envisage

a set of one-to-one mapped population records (i.e.  $U$ ) conceptually. Similarly for other population enumerations, such as  $A$  by the administrative sources. After all, when linking or comparing the different enumeration results, one is actually dealing with records (rather than persons) in different sets.

It follows that, while any imputed record taken from  $C$  *always* causes an enumeration error since  $C \cap (U \setminus C) \equiv \emptyset$  by definition, enumeration errors are avoided whenever a record is taken from  $A \cap (U \setminus C)$ . Thus, regardless the respective assumptions, imputing  $U_{imp}$  by administrative records in  $A \setminus C$  has always the advantage that it may greatly reduce the resulting enumeration errors of  $U_{imp}$  compared to imputing from the census records  $C$ .

It should be noted that imputing census enumerated records may well be acceptable for aggregate statistics, in which respect the accuracy of individual records depends on having the right kind of imputed characteristics rather than the avoidance of enumeration error *per se*. However, as administrative data and data linkage at scale are now commonplace and used extensively to provide statistics or data for analysis, it matters more if the imputed records are real and linkable, whereas the items that are only collected in the census can be imputed from the enumerated census records *given* such imputed administrative records. This would enable uses of the imputed census database beyond what was previously possible, as the imputed records will not have to be excluded and adjusting analyses of linked data for census unit or item nonresponse will be a less demanding issue.

In this paper, we shall address the aforementioned gap in the literature by developing some techniques for estimating the enumeration errors of imputing administrative records for census missing enumerations. As the imputation ideally should only use the records  $A \setminus C$ , a particular difficulty we need to overcome is the *linkage errors* that exist whenever one does not have a unique identity key that can be used to link  $C$  and  $A$  unequivocally.

Notice that to focus on imputation for the census missing enumerations by the records in  $A$ , we may assume tacitly in the theoretical development below that  $C$  is free of erroneous records, i.e.

$$C \setminus U = \emptyset \tag{1}$$

This is of course reasonable if census erroneous enumerations are negligible. However, even when this is not the case, any erroneous records in  $C$  would not matter to the enumeration errors induced by imputation from  $A$ , the purpose of which is to deal with census missing enumerations, except when such a record is falsely unlinked to  $A$  and the true matching record in  $A$  is later selected as an imputed record. The extent of such effects is limited because it requires all three events to take place: erroneous enumeration by  $C$ , false negative link to  $A$ , selection for imputation among the unlinked records in  $A$ . Whereas to account for census erroneous enumerations, it is common to estimate their total by a coverage survey and adjust the population estimate accordingly, and one can extend this survey to cover  $U_{imp}$  as well.

Now, given (1), Figure 1 illustrates how  $U$ ,  $C$  and  $A$  are related to each other when the records in  $C$  and  $A$  cannot be matched to each other in an error-free manner. Let  $C_L$  be the set of records in  $C$  that can be linked to  $A$  *uniquely*

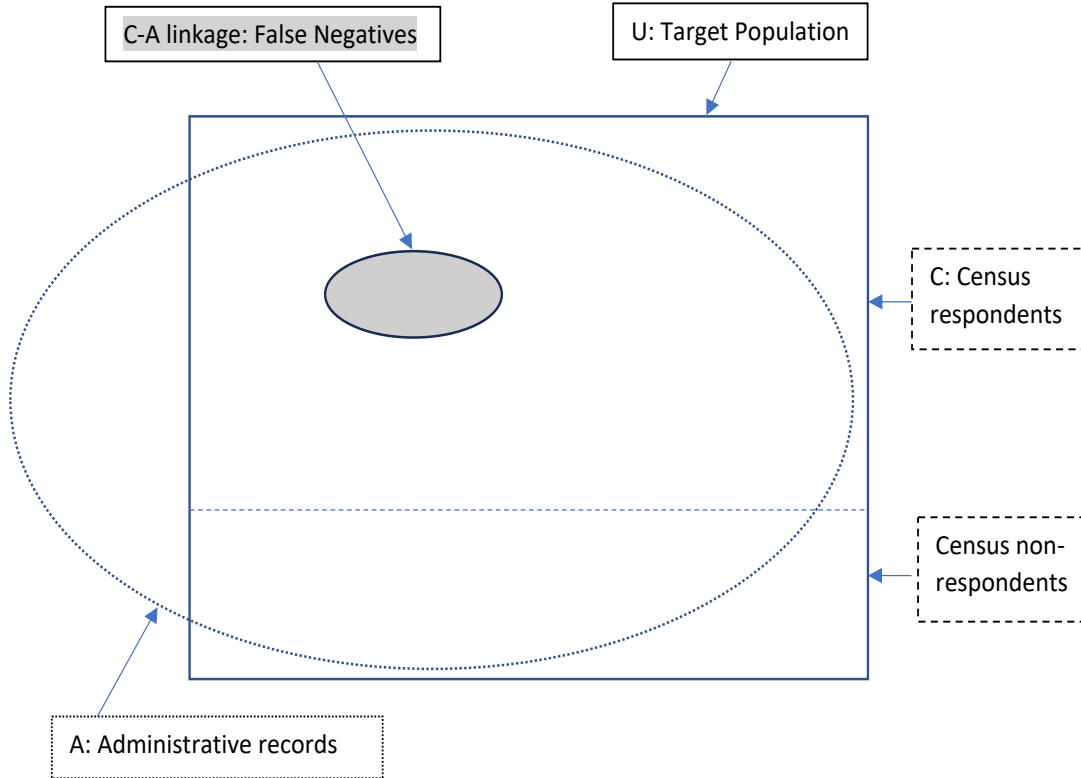


Figure 1: Target population records ( $U$ , solid square), census respondents ( $C$ , dashed separation from census non-respondents), administrative records ( $A$ , dotted oval), false negatives of linkage between  $C$  and  $A$  (solid shaded oval)

according to some given record linkage procedure, such that  $C_{NL} = C \setminus C_L$  contains the records in  $C$  with *negative* linkage outcomes, i.e. the unlinked census records. In the presence of linkage errors  $C_{NL}$  would consist of both the *false negative (FN)* records  $C_{FN}$  and the *true negative (TN)* records  $C_{TN}$ , where a record in  $C_{NL}$  belongs to  $C_{FN}$  if it actually has a matching record in  $A$  or it belongs to  $C_{TN}$  if it does not have a matching record in  $A$  at all. Notice that  $C_{FN}$  is marked by the small shaded oval in Figure 1 and  $C_{TN}$  are the “census respondents” outside the large oval, where record linkage between  $C$  and  $A$  reveals only their union (as  $C_{NL}$ ) but not each set on its own.

It is then necessary to estimate the FN and TN numbers among  $C_{NL}$ , in order to assess the enumeration errors when  $U_{imp} \subset A \setminus C_L$ . This is because

$$A_{FN} \cap U_{imp} \neq \emptyset \quad (2)$$

can occur, where  $A_{FN}$  are the records in  $A$  that should have been linked to  $C_{FN}$ , which means that a population record in  $U$  would belong to both  $C$  and  $U_{imp}$ . Moreover, it is necessary to estimate the number of true matches between  $U$  and  $A$ , called the in-scope records in  $A$ . This is because

$$U_{imp} \setminus U \neq \emptyset \quad (3)$$

can occur if  $U_{imp}$  contains some of the rest out-of-scope (or erroneous) records in  $A$  (i.e. signified by the part of large oval outside the square in Figure 1). This

would lead to a record which does not belong to  $U$  to appear in  $U_{imp}$ .

Thus, we need to consider two types of enumeration error (2) and (3) that can arise when  $U_{imp} \subset A \setminus C_L$ . The other records in  $A \setminus C_L$  (i.e. the overlap of the large oval and the “census non-respondents” in Figure 1) do not cause enumeration errors if they are included in  $U_{imp}$ ; indeed, making use of these records is the primary motivation for imputing administrative records.

It should be immediately pointed out that the delineation of enumeration error above has been simplified by making an assumption about the linkage between  $C$  and  $A$ , i.e. all the records in  $C_L$  are correctly matched to  $A$  such that there are no *false positive (FP)* linkage outcomes, denoted by

$$C_{FP} = \emptyset \tag{4}$$

A false positive outcome arises when two records in  $C$  and  $A$  are classified as linked when they actually refer to different individuals. This could cause enumeration errors induced by imputation, if the linked administrative record actually corresponds to a census missing enumeration, since this record would then have no chance of being included in  $U_{imp} \subset A \setminus C_L$ . However, since a false positive outcome would require the key variables of two un-matched records to agree in a very peculiar way (so as to be linked by mistake), the probability of which is generally negligible compared to a false negative outcome (which could be caused by as few as a single key variable perturbation). Thus, even though the assumption (4) may not be exactly true, it can be safely adopted for practical linkage error adjustments (e.g. Ding and Fienberg, 1994), as long as record linkage makes use of a sufficient number of key variables. While the approach here has been to eliminate False Positives, interesting further work is to incorporate them into this framework. This would lead to an explicit trade-off between enumeration errors due to False Positives and False Negatives. This could be necessary in situations where there is an insufficient number of discriminating linking variables.

In what follows, we shall first set out in Section 2 the estimation methods for enumeration errors induced by imputing administrative records for census missing enumerations. The methods will be illustrated by an application to the Australian Census 2016 in Section 3. Finally, Section 4 contains some closing remarks and topics for future research.

## 2 Estimation of enumeration error

Given any record in  $C \cup A \cup U$ , let  $\delta = 1, 0$  indicate whether it belongs to  $C$  or not, let  $a = 1, 0$  indicate whether it belongs to  $A$  or not, and let  $u = 1, 0$  indicate whether it belongs to  $U$  or not. A record in  $C$  (or  $A$ ) has a matching record in  $A$  (or  $C$ ) iff  $a\delta = 1$ , and a record in  $U$  (or  $A$ ) has a matching record in  $A$  (or  $U$ ) iff  $au = 1$ . Moreover, let  $r = 1, 0$  indicate whether a given record in  $C$  or  $A$  is linked by record linkage between  $C$  and  $A$ . In particular, any record belongs to  $C_L$  iff  $r\delta = 1$ , whereas it belongs to  $C_{NL}$  iff  $(1 - r)\delta = 1$ .

For the estimation theory to be developed below, we shall treat  $U$  and  $A$  as fixed, which means that  $u$  and  $a$  are constants whether or not they are observed.

The indicator  $\delta$  will be treated as a random variable associated with census enumeration. In particular, let  $y$  be a feature vector that renders homogeneous census response probabilities, denoted by

$$\pi(y) = \begin{cases} \Pr(\delta = 1 \mid u = 1, a, y) = \Pr(\delta = 1 \mid u = 1, y) \\ \Pr(\delta = 1 \mid u = 0, a, y) = 0 \end{cases} \quad (5)$$

where the second nil probability follows from (1), and the first *post-stratification* probability is standard in the practice of census coverage error adjustment (e.g. Wolter, 1986) given the judicious choice of  $y$ . In particular,  $y$  consists of geography (state of Australia), age and sex in Section 3 later.

Moreover, the indicator  $r = 1$  will be treated as a random variable associated with record linkage between  $C$  and  $A$ , i.e. conditional on census enumeration. In particular, let  $z$  be the vector indicators for the presence of the key variables for linkage, such that the FN probability is given by

$$\lambda(z, y) = \Pr(r = 0 \mid a\delta = 1, z, y) = 1 - \Pr(r = 1 \mid a\delta = 1, z, y) \quad (6)$$

since the linkage outcomes are completely determined by how the key variables of two records compare to each other. In practice, the more key variables are available in  $z$  the smaller  $\lambda$ -probability can be expected. In particular, for the application later in Section 3,  $z$  consists of indicators for the availability of SA1 (small area geography of Australia), address, date of birth, first name, second name and 5-year-ago SA1.

## 2.1 Estimation related to matches between $C$ and $A$

Since all the records in  $C_L$  are correctly matched to  $A$  by (4), we have  $r = 1$  only if  $a = 1$ . This yields

$$N(r\delta = 1, z, y) = N(a = 1, r\delta = 1, z, y)$$

where  $N(\cdot)$  denotes the number of records specified inside the parentheses. Under the model (6) of FN probability, we have then

$$E\{N(r\delta = 1, z, y)\} = \Pr(r = 1 \mid a\delta = 1, z, y) E\{N(a\delta = 1, z, y)\} \quad (7)$$

where  $N(a\delta = 1, z, y)$  is the corresponding number of true matches between  $C$  and  $A$ . To estimate the probability in (7), we assume further that

$$N(a\delta = 1, z, y, \alpha = 0) = N(\delta = 1, z, y, \alpha = 0) \quad (8)$$

where  $\alpha = 0$  if a census record belongs to a native born person and 1 otherwise, and  $\alpha$  is observed for all the records in  $C$ . This is a reasonable assumption in any country, provided the underlying administrative sources of  $A$  include birth registration, school education, taxation and health care, such as Australia. However, note that the general idea is to choose appropriately a subset of the population for the identifying equation (8). For example, Statistics New Zealand (2019c) assumes that adults born in NZ with a taxable income will have a

record on the administrative data source. Note also that here we only assume a matching record exists in  $A$  for a native born person enumerated in  $C$ , not that one would always manage to link the two records.

Combining (7) and (8), we obtain

$$\Pr(r = 1 \mid a\delta = 1, z, y, \alpha = 0) = \frac{E\{N(r\delta = 1, z, y, \alpha = 0)\}}{E\{N(\delta = 1, z, y, \alpha = 0)\}}$$

where  $(r, z, y, \alpha)$  are observed for all the census records  $C$ . We can therefore estimate  $\Pr(r = 1 \mid a\delta = 1, z, y, \alpha = 0)$  given the observed counts on the right-hand side over  $C_L$  and  $C$ , respectively, either directly or using any suitable models such as logistic regression to reduce the variance of estimation.

The estimated FN probability  $\Pr(r = 0 \mid a\delta = 1, z, y, \alpha = 0)$  can be applied as the estimates of  $\Pr(r = 0 \mid a\delta = 1, z, y)$  regardless  $\alpha$  by virtue of (6). For instance, by (7), we can obtain the estimates of total matches  $N(a\delta = 1, z, y)$  between  $C$  and  $A$ , by expanding each observed count in  $C_L$  by the inverse of the estimated probability  $\Pr(r = 1 \mid a\delta = 1, z, y)$ .

To summarise, we are able to estimate the FN probabilities and the number of true matches between  $C$  and  $A$  by  $(z, y)$  using the following assumptions concerning record linkage between  $C$  and  $A$ .

- No false positive linkage outcomes (4).
- The false negative probability (6) is a constant given  $(z, y)$ .
- Matches exist between  $C$  and  $A$  for any native born persons in  $C$ , i.e. (8).

Notice that the condition that  $\alpha = 0, 1$  is always observed in census may be introduced as a criterion for a valid census record if necessary. In addition, we have used the observed counts over  $C$  and  $C_L$  directly given the assumption (1), such as  $N(\delta = 1, z, y, \alpha = 0)$ . In case census erroneous enumerations are not negligible, one would need to subtract their estimated totals (e.g. by a coverage survey) from the counts before estimating the FN probabilities.

## 2.2 Estimation related to matches between $U$ and $A$

As explained in Section 1, the FN probabilities considered above are related to the enumeration error (2), whereas the other enumeration error (3) occurs if an erroneous record in  $A$  is included in  $U_{imp}$ . Let the likelihood of an erroneous record in  $A$  depend on a feature vector  $x$  of the various patterns and strengths of signs-of-life in the underlying administrative sources. In particular, in the application in Section 3,  $x$  is made up of activity on Medicare, Social Security, and Personal Income Tax (e.g. 1, 6, 12 months), death status, migration status, age and geography (state of Australia).

The relevant numbers of in-scope records (i.e. matches between  $U$  and  $A$ ) in the different parts of  $A$  can then be specified as  $N(au = 1, y, x)$ . These matches are related to those between  $C$  and  $A$  by

$$\begin{aligned} E\{N(au\delta = 1, y, x)\} &= \Pr(\delta = 1 \mid au = 1, y, x) N(au = 1, y, x) \\ &= \Pr(\delta = 1 \mid u = 1, y) N(au = 1, y, x) \end{aligned}$$

where the first equality follows from the assumption (1) and the second equality from (5). Provided error-free matching of  $C$  and  $A$ , one would observe directly

$$N(au\delta = 1, y, x) = N(a\delta = 1, y, x)$$

such that  $N(au = 1, y, x)$  can be estimated given an estimate of the census response probability  $\Pr(\delta = 1 \mid u = 1, y)$ . Whereas in the presence of FN linkage outcomes between  $C$  and  $A$ , one would need to estimate  $N(a\delta = 1, y, x)$  via (7), where  $N(a\delta = 1, y, x) = \sum_z N(a\delta = 1, z, y, x)$ , i.e.

$$E\{N(a\delta = 1, y, x)\} = \sum_z \frac{E\{N(r\delta = 1, z, y, x)\}}{\Pr(r = 1 \mid a\delta = 1, z, y)} \quad (9)$$

In other words,  $N(au = 1, y, x)$  can be estimated given explicitly the census response probabilities (5), and the FN probabilities estimated in Section 2.1 are needed additionally given linkage error between  $C$  and  $A$ .

The census response probabilities (5) can be estimated in different ways. First, where census under-coverage survey is carried out in connection with the census, these probabilities can be estimated using the so-called dual system estimation methodology (e.g. Hogan, 1993; Brown, et al., 1999) or another suitable method (e.g. Chipperfield et al., 2017). Next, without implementing any under-coverage survey, it may still be possible to estimate these probabilities by linking  $C$  and a so-called trimmed administrative dataset with negligible erroneous records, see Dunne and Zhang (2023) for an example from Ireland in the absence of linkage errors, or Chipperfield et al. (2024) for an approach in the presence of linkage errors.

Now, to estimate the enumeration error (3), where  $U_{imp} \subset A \setminus C_L$ , we only need to derive the number of matches between  $U$  and  $A$  among the unlinked records in  $A$ , i.e.  $N(r = 0, au = 1, y, x)$  instead of the overall number of matches  $N(au = 1, y, x)$ . This can be given as

$$\begin{aligned} N(r = 0, au = 1, y, x) &= N(au = 1, y, x) - N(r\delta = 1, au = 1, y, x) \\ &= N(au = 1, y, x) - N(r\delta = 1, y, x) \end{aligned}$$

where the first equality follows by definition, and the second equality follows from (1) and (4), i.e.  $\delta = 1$  only if  $u = 1$  and  $r = 1$  only if  $a = 1$ , and  $N(r\delta = 1, y, x)$  is directly observed over the linked census records  $C_L$ .

Given  $N(r = 0, au = 1, y, x)$ , we can calculate the proportion of in-scope records in  $A$  given  $(y, x)$  and that the records are unlinked ( $r = 0$ ) as

$$\gamma(y, x) = \frac{N(r = 0, au = 1, y, x)}{N(r = 0, a = 1, y, x)} \quad (10)$$

where the denominator  $N(r = 0, a = 1, y, x)$  is an observed count over  $A \setminus C_L$  following record linkage between  $C$  and  $A$ .

We could also calculate the overall proportion of in-scope records in  $A$  given

$(y, x)$  regardless whether the records are linked to  $C$ ,

$$\gamma_1(y, x) = \frac{N(au = 1, y, x)}{N(a = 1, y, x)}, \quad (11)$$

where the denominator is known and the numerator is given by (9). This would be useful outside the census occasions.

### 2.3 Robust estimation of $\gamma(y, x)$ and $\gamma_1(y, x)$

The approach of Section 2.2 to  $\gamma(y, x)$  may be sensitive to small or empty cell counts, since the estimate of  $N(au = 1, y, x)$  via (9) is a sum over the three-way breakdown of  $C_L$  by  $(z, y, x)$ . We now develop an alternative estimator of  $\gamma(y, x)$ , which is approximate in theory but more robust in practice, since it requires only two-way breakdowns of  $C_L$  by  $(z, y)$  and  $(y, x)$ , respectively.

We start by noticing that

$$\begin{cases} N(r = 0, au = 1, y, x) = N(r = 0, au\delta = 1, y, x) + N(\delta = 0, au = 1, y, x) \\ N(\delta = 1, au = 1, y, x) = N(r\delta = 1, y, x) + N(r = 0, au\delta = 1, y, x) \end{cases}$$

where  $N(r = 0, au\delta = 1, y, x)$  over the FN records is much smaller than the other terms. For instance, in our application to the Australian data later, the ratio between  $N(\delta = 0, au = 1, y, x)$  and  $N(r = 0, au\delta = 1, y, x)$  is about 13 to 1, and that between  $N(r\delta = 1, y, x)$  and  $N(r = 0, au\delta = 1, y, x)$  is about 25 to 1. We therefore adopt the following approximations

$$\frac{N(r = 0, au = 1, y, x)}{N(r = 0, au = 1, y)} \approx \frac{N(\delta = 0, au = 1, y, x)}{N(\delta = 0, au = 1, y)} \quad (12a)$$

$$\frac{N(\delta = 1, au = 1, y, x)}{N(\delta = 1, au = 1, y)} \approx \frac{N(r\delta = 1, y, x)}{N(r\delta = 1, y)} \quad (12b)$$

by taking the leading term but omitting the contributions from the FN records. Moreover, we can combine the two approximations by applying (5) twice,

$$\begin{aligned} \frac{E\{N(\delta = 0, au = 1, y, x)\}}{E\{N(\delta = 0, au = 1, y)\}} &= \frac{\{1 - \Pr(\delta = 1 \mid u = 1, y)\} N(au = 1, y, x)}{\{1 - \Pr(\delta = 1 \mid u = 1, y)\} N(au = 1, y)} \\ &= \frac{\Pr(\delta = 1 \mid u = 1, y) N(au = 1, y, x)}{\Pr(\delta = 1 \mid u = 1, y) N(au = 1, y)} \\ &= \frac{E\{N(\delta = 1, au = 1, y, x)\}}{E\{N(\delta = 1, au = 1, y)\}} \end{aligned}$$

i.e. the ratio among the census non-respondents is equal to that among the census respondents given homogenous census response probabilities (5).

Meanwhile, the denominator on the left-side of the first approximation (12a) is given by

$$N(r = 0, au = 1, y) = N(au = 1, y) - N(r\delta = 1, au = 1, y)$$

The second term on the right side is directly observed as  $N(r\delta = 1, y)$  over  $C_L$ . For the first term, we have

$$\frac{N(au = 1, y)}{N(u = 1, y)} = \frac{N(au = 1, y) \Pr(\delta = 1 \mid au = 1, y)}{N(u = 1, y) \Pr(\delta = 1 \mid u = 1, y)} = \frac{E\{N(au\delta = 1, y)\}}{E\{N(u\delta = 1, y)\}}$$

since the census response probability (5) does not depend on  $a$  given  $y$ . Due to (1), we observe  $N(u\delta = 1, y)$  directly as  $N(\delta = 1, y)$  in  $C$ , whilst  $N(au\delta = 1, y)$  is equal to  $N(a\delta = 1, y)$ , i.e. the number of matches between  $C$  and  $A$  given  $y$ , which can now be estimated as the sum of  $N(a\delta = 1, z, y)$  over  $z$  using (7).

Taking together all the development above, we obtain

$$E\{N(r = 0, au = 1, y, x)\} \approx E\{N(r = 0, au = 1, y)\} \frac{E\{N(r\delta = 1, y, x)\}}{E\{N(r\delta = 1, y)\}} \quad (13)$$

where

$$E\{N(r = 0, au = 1, y)\} = N(u = 1, y) \frac{E\{N(a\delta = 1, y)\}}{E\{N(\delta = 1, y)\}} - E\{N(r\delta = 1, y)\} \quad (14)$$

and an estimate of  $N(u = 1, y)$  is available given the estimated census response probability (5). This yields an approximate estimator of  $\gamma(y, x)$  given by (10), where the two-way counts of  $(y, x)$  and  $(z, y)$  in  $C_L$  are needed to apply (13) and (14), respectively, but the three-way counts of  $(z, y, x)$  in  $C_L$  are no longer necessary. To calculate  $\gamma_1(y, x)$  we note that its denominator in (11) is known. The numerator  $N(au = 1, y, x)$  is the sum of  $N(r = 1, au = 1, y, x)$  that is known and  $N(r = 0, au = 1, y, x)$  that is given by (13).

In the application below,  $\gamma_1$  will be used to select the records in  $A$  for census imputation, in comparison to a rule-based criterion that is typical otherwise. We note that any arbitrary selection criterion could have been applied to illustrate the methods developed here for estimating the enumeration errors. However, as will be explained, regardless the imputation rule, we always use  $\gamma$  to measure the enumeration error associated with imputing an *unlinked* record from  $A$ , because the records linked to  $C$  do not cause enumeration errors.

### 3 Application to Australian Census 2016

We illustrate the estimation methods developed above by an application in the context of Australian Census 2016.

#### 3.1 Set-up

The official number of residents is 24.2 million at the time of the 2016 Census (Australian Bureau of Statistics, 2016), referred to as Population Total and denoted by  $N_U$ . The Census set  $C$  contains 22.2 million records belonging to responding residents (Australian Bureau of Statistics, 2017). The number of redundant records in  $C$  is 0.27 million, due to duplicated enumeration, which will be counted as the only source of Census over-coverage error. Under-counting of  $N_U$  by  $C$  due to the missing enumerations is then 2.27 million, of which 0.7

million are temporarily overseas at the time. Although these overseas persons could not have been enumerated in  $C$ , we count them together with the other census non-respondents to simplify the exposition here.

Next, the administrative data source is the Multi-Agency Data Integration Project (MADIP) in Australia (ABS 2020a). MADIP is a secure data asset with legislative and privacy protections, along with policies and standards, and safe data handling practices. It combines information on health, education, government payments, income and taxation, employment, population demographics (age and sex) and small area geography. It includes records from the Medicare Consumer Directory, DOMINO Centrelink Administrative Data and Personal Income Tax. It aims to cover the “ever-resident” population, including people who have permanently emigrated or who have died. It is expected that the MADIP has a high overlap with Australian residents. Some possible exceptions include: recent migrants who, depending upon their visa type, may be ineligible for government services; international students; Aboriginal and Torres Strait Islander peoples, also noting that identification of these peoples by the Census and MADIP may be inconsistent; people experiencing homelessness, and infants (due to processing delays). However, there is no data item available for all the records on MADIP to indicate that it belongs to a resident (i.e. such a data item is available for some but not all records).

Let  $A$  be derived from the MADIP, which consists over 33 million records. Of the records that are not linked to  $C$  above,  $A \setminus C_L$ , let  $U_{imp}$  contain those whose associated score  $1 - \gamma_1(y, x)$  is below a threshold value  $p$ ,  $0 < p < 1$ , which can be thought of as ‘likely residents who cannot be verified by the Census’. Although it is possible to render the amount of erroneous records in  $U_{imp}$  negligible by choosing a sufficiently low threshold  $p$ , which however would aggravate the under-coverage error at the same time, a less extreme choice of  $p$  may entail a better balance. Some different choices of  $p$  will be compared later in Section 3.3. However, notice that our aim here is not to propose the best census imputation method, but to illustrate the methods for estimating the enumeration errors of census imputation thereby enabling such comparisons at all.

We consider two ways of combining  $U_{imp}$  with the census records to yield the imputed census set. First, we refer to the imPuted Link Set (PLS) as

$$C_L^* = C_L \cup U_{imp}$$

i.e. combining  $U_{imp}$  with the linked records between  $C$  and  $A$ . Since any FN record cannot appear more than once in  $C_L^*$ , all erroneous enumerations by  $C_L^*$  are due to  $U_{imp}$ , which can be estimated as

$$N_{err}(C_L^*) = N_{err}(U_{imp}) = \sum_{i \in U_{imp}} 1 - \gamma(y_i, x_i)$$

where  $\gamma(y, x)$  is defined by (10). Moreover, denote by  $N_L$  the number of records in  $C_L$ , the total missing enumerations of  $C_L^*$  can be estimated as

$$N_{mis}(C_L^*) = N_U - N_L - \sum_{i \in U_{imp}} \gamma(y_i, x_i)$$

Next, combining  $U_{imp}$  with  $C$  including  $C_{NL}$  as well, we refer to the Census-Imputed set analogously to  $C_L^* = C_L \cup U_{imp}$  above as

$$C^* = C \cup U_{imp}$$

On the one hand, the FN records can now possibly appear twice in  $C^*$ , which is a type of enumeration error absent in  $C_L^*$ ; on the other hand, the in-scope TN records in  $C$  are now included in  $C^*$ , which are missing from  $C_L^*$ . An estimate of the number of erroneous enumerations by  $C^*$  is given as

$$N_{err}(C^*) = N_{err}(C) + N_{err}(U_{imp}) + \sum_{i \in U_{imp}} \pi(y_i) \gamma(y_i, x_i) > N_{err}(C_L^*)$$

where  $N_{err}(C)$  is the number of duplicated records in  $C$ , and  $N_{err}(U_{imp})$  is just the number of erroneous records in  $C_L^*$  as explained above, and the last term refers to the double-counted records in  $C_{FN}$  which requires the matching record in  $U_{imp} \subset A \setminus C_L$  to have  $\delta = 1$  in addition to  $u = 1$  given  $(r, a) = (0, 1)$ . Whereas an estimate of the number of missing enumerations by  $C^*$  is given as

$$N_{mis}(C^*) = N_{mis}(C_L^*) - N_{TN} < N_{mis}(C_L^*)$$

where  $N_{TN}$  is the estimated total of TN records in  $C_{NL}$  given by

$$N_{TN} = \sum_{z,y} N(r = 0, \delta = 1, z, y) - \lambda(z, y) N(a\delta = 1, z, y)$$

via the FN probability  $\lambda(z, y)$  and match count  $N(a\delta = 1, z, y)$  between  $C$  and  $A$ .

### 3.2 Results: FN and TN by Census-MADIP linkage

The estimated TN and FN links from the Census-MADIP (or  $C$ - $A$ ) linkage are given in Table 1, relatively to the total Census enumeration. If the unlinked Census records are excluded from the imputed census database, then the TN links among them would cause under-coverage error since it is not possible for them to be part of the imputed records selected from the MADIP. Whereas if the unlinked Census records are included in the imputed census database, then the FN links among them could cause double-counting (or duplication) between Census and the records taken from the MADIP.

Not only is the FN linkage error clearly greater than the TN linkage error in all the results here, the former also varies much more across the population. The proportion of FN links is higher in less densely populated areas and varies notably by Indigenous status. In particular, at the national level, the FN error for Aboriginal and Torres Strait Islanders is about three times that of the other people (11% vs 3.6%); whereas in the Northern Territory, the state with the highest proportion of Aboriginal and Torres Strait Islanders, the proportion of FN links is as high as 66% in Very Remote areas.

The assumption (8) is particularly useful in cases where the proportion of unlinked census records is relatively high, such as in the outer regional or remote areas of Australia, where it leads to the finding that FN linkage is much

Table 1: Census-MADIP TN and FN links among Census enumeration

	<b>True Negative (TN, %)</b>	<b>False Negative (FN, %)</b>
National	1.1	4.0
<b>By Indigenous Status</b>		
Non-Indigenous	1.1	3.6
Indigenous	1.1	11
<b>By Area Type</b>		
Major cities	0.5	3.9
Inner regional Australia	0.6	4.5
Outer regional Australia	0.1	7.4
Remote Australia	2.0	19
Very remote Australia	0.5	22
<b>In Northern Territory</b>		
Major cities	0.7	4.6
Inner regional Australia	0.6	6.2
Outer regional Australia	1.7	15
Remote Australia	2.4	40
Very remote Australia	2.2	66

more likely than TN linkage in those cases, which matters to the evaluation of the coverage error that would result from different census imputation schemes. We notice also that Chipperfield et al. (2024) apply the relevant estimates of FN links in a study of triple-system estimation of population size, involving census, census coverage survey and the MADIP, the results of which provide additional corroboration to the estimation method built on (8).

### 3.3 Imputation of unlinked MADIP records

Next, the two ways of imputing for the census we consider here both require a set of records  $U_{imp}$  to be selected from the unlinked MADIP records  $A \setminus C_L$ . We shall refer to any selection criterion as a ‘scoping method’. In particular, let a rule-based criterion require some relevant sign-of-life activity in the last 24 months; whereas for a (likelihood) threshold method, let  $U_{imp}$  include any record  $i$  if  $1 - \gamma_1(y_i, x_i) < p$  and a chosen threshold value  $p$ .

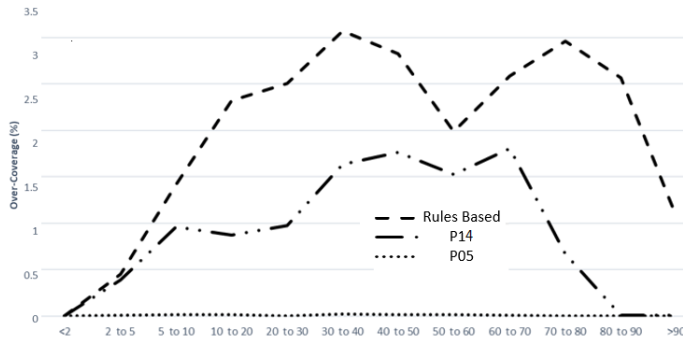


Figure 2: PLS erroneous enumerations by scoping method

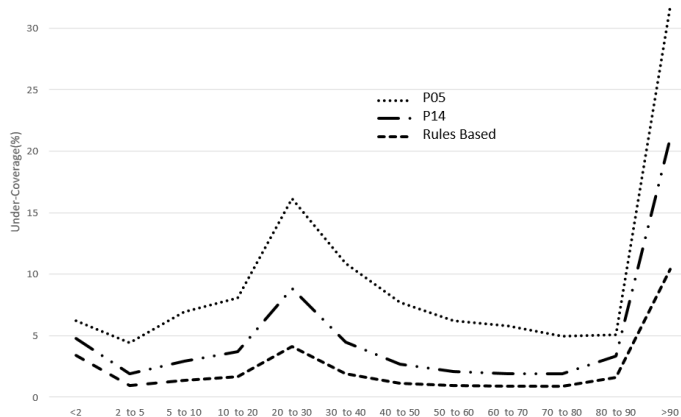


Figure 3: PLS missing enumerations by scoping method

Figures 2 and 3 compare the rule-based method and two threshold methods with, respectively,  $p = 0.5\%$  (P05 in short) and  $p = 14\%$  (P14 in short). First, the estimated proportions of erroneous enumeration in  $U_{imp}$  (or  $C_L \cup U_{imp}$ ) based on (10) are given in Figure 2 for the different age groups: by adopting a sufficiently low threshold for being out-of-scope among the unlinked MADIP records, the P05-method can practically eliminate erroneous enumeration, while the rule-based method yields the highest erroneous enumeration error in all the age groups (apart from those less than 24 months old). Next, Figure 3 shows for all the age groups the estimated under-coverage rate of the records  $C_L \cup U_{imp}$  resulting from the three scoping methods: the P05 method leads to the highest under-coverage rate in all the age groups (apart from those at the two ends of the age spectrum), otherwise the rule-based method and the P15-method are quite similar in terms of the resulting under-coverage rates.

This illustrates how the choice of the imputed records  $U_{imp}$  can cause varying enumeration errors, when these are taken from the administrative sources. The estimation method developed in this paper are useful for assessing these enumeration errors and the resulting coverage errors of the imputed census. Below we examine the Census-Imputed records  $C \cup U_{imp}$  and the PLS records  $C_L \cup U_{imp}$  where  $U_{imp}$  is given by the P14-method, in comparison to the Census records  $C$  without imputation.

### 3.4 Results: Enumeration and coverage errors

The number of records in Census, Census-Imputed and PLS are 22.2, 24.7, and 23.9 million, respectively. Compared to the official Population Total of 24.2 million, Census-Imputed has a positive net coverage error, while Census and PLS have negative net errors. Figure 4 shows all the counts by one-year groups of age. The Census counts are noticeably below the official counts for the under 70 year olds. This gap is significantly reduced by Census-Imputed due to the additional 2.5 million records  $U_{imp}$ . In fact, the Census-Imputed counts are slightly higher than the official counts in much of the under 70 year old age range. The PLS counts are often close to the Official counts, with the notable exception for the 15-35 years olds.

For each given set of enumeration or imputed records, the net coverage error

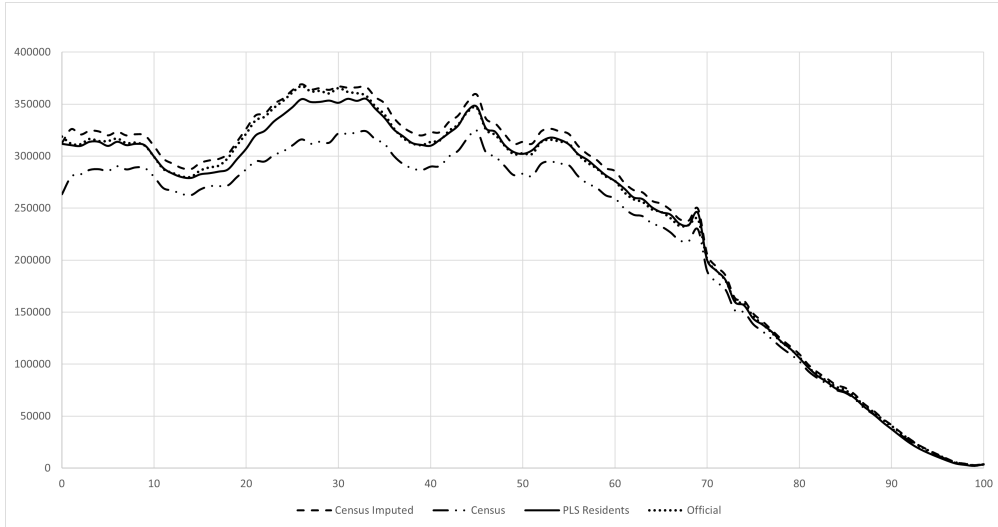


Figure 4: Age counts by Census, Census-Imputed or PLS

Table 2: Enumeration and coverage errors, all in millions. Census 2016 ( $C$ ), imputed MADIP records ( $U_{imp}$ ), Census-MADIP links ( $C_L$ ).

	$C$	$C \cup U_{imp}$	$C_L \cup U_{imp}$
Duplicated enumeration in $C$	0.3	0.3	N/A
+ Erroneous enumeration in $U_{imp}$	N/A	0.3	0.3
+ Double-counting by $C$ and $U_{imp}$	N/A	0.2	N/A
= Over-coverage (E)	0.3	0.7	0.3
Missing enumeration by $C$	2.3	N/A	N/A
Missing enumeration by $C_L \cup U_{imp}$	N/A	0.5	0.5
– TN of Census-MADIP linkage in $C \setminus C_L$	N/A	0.2	N/A
= Under-coverage (M)	2.3	0.3	0.5
Net coverage error (E – M)	-2.0	0.4	-0.2
Gross coverage error (E + M)	2.5	1.0	0.8

(shown in Figure 4) is the difference between over- and under-coverage errors. To complete the picture, one must also take into account the gross coverage error given as the sum of over- and under-coverage errors; moreover, either the over- or under-coverage error needs to be decomposed into the relevant enumeration errors. Table 2 shows the enumeration errors and the resulting coverage errors of Census, Census-Imputed and PLS, respectively.

First, as mentioned before, duplicated enumeration records will be treated as the over-coverage error of the Australian Census 2016 here, and missing enumeration is the only cause of under-coverage error. There are 0.27 million redundant records in Census, i.e. there are  $22.2 - 0.27$  million distinct enumeration records in Census and any one of the 0.27 million redundant records is a duplicate of one of these distinct records, causing a duplicated enumeration error. The Census net and gross errors follow as given in Table 2.

Take now PLS given as  $C_L^* = C_L \cup U_{imp}$ . None of the redundant Census records can be included in  $C_L$ , because Census-MADIP links are only acceptable for the distinct records in Census — hence, duplicated enumeration error

is absent. Next,  $N_{err}(C_L^*) = N_{err}(U_{imp}) = 0.3$  million is obtained as explained in Section 3.1, due to the erroneous enumeration error of  $U_{imp}$ . Since the unlinked Census records are excluded from PLS, double-counting between them and the imputed MADIP records is absent. This yields 0.3 million as the PLS over-coverage error. Moreover, as explained in Section 3.1,  $N_{mis}(C_L^*) = 0.5$  million is obtained as the missing enumeration error of PLS, which is also its under-coverage error. It can be seen that both the PLS net and gross coverage errors are much reduced compared to Census without imputation.

Given the enumeration errors of Census and PLS, we only need to estimate two more coverage error components for Census-Imputed  $C^* = C \cup U_{imp}$ . The first of them is double-counting between  $C$  and  $U_{imp}$ , because a falsely unlinked in-scope record in the MADIP can be both selected into  $U_{imp}$  and captured by Census. This is the last component of  $N_{err}(C^*)$ , in addition to  $N_{err}(C) = 0.27$  million and  $N_{err}(U_{imp}) = 0.3$  million already accounted for above, which can be estimated as described in Section 3.1. The last coverage error component for Census-Imputed is the number of TN links (denoted by  $N_{TN}$  in Section 3.1) among the distinct unlinked Census records (outside the 0.27 million redundant records that are also unlinked). The estimation of the total TN links has been discussed in Section 3.2. One needs to subtract  $N_{TN}$  from the missing enumeration count by  $C_L \cup U_{imp}$  to obtain the missing enumeration count by  $C \cup U_{imp}$ , because the latter includes all these TN links.

As can be seen in Table 2, Census-Imputed has both a lower under-coverage error and a higher over-coverage error than PLS, as explained in Section 3.1. Both the net and gross coverage errors are somewhat reduced with PLS than Census-Imputed. However, it is possible to remove the duplicated records in Census, which would improve Census-Imputed without affecting PLS. (While it is easier to identify duplicates at the same address, collecting multiple recent addresses for Census records helps to identify duplicates at different addresses by, for example, by linking the Census to itself.) Only the Census-MADIP linkage errors would then matter to the choice whether or not to include the unlinked Census records. Had this linkage been perfect, such that all the links are true matches and all the true matches are linked, double-counting by  $C$  and  $U_{imp}$  would not have been possible. Since it is obviously beneficial to include the unlinked but in-scope Census records, one would only need to concentrate on the trade-off between erroneous and missing enumerations by the choice of imputed administrative records  $U_{imp}$ .

### 3.5 A thought experiment

This paper was partly motivated by the critical need for contingency planning in the potential event of a low response rate for the Australian 2021 Census due to the Covid pandemic. In the event of such a contingency, the Australian Bureau of Statistics obtained agreement from custodial agencies to use MADIP records for unit imputation of Census 2021 nonresponse. In the thought experiment described below, we evaluate the enumeration and coverage errors associated with census imputation in a simulated low Census response scenario.

To simulate a low response scenario, 1.5 million dwellings enumerated in the 2016 Census are randomly removed in the specific subpopulations that

are more prone to increasing nonresponse:

- 75,000 Indigenous dwellings,
- 75,000 Secure apartment block dwellings,
- 100,000 Dwellings with a recent migrant (within last 3 years),
- 150,000 Younger aged dwellings (only of persons 15-29),
- 200,000 Dwellings that required high follow-up (more than 2 visits),
- 400,000 Dwellings in drop-off areas,
- 500,000 Dwellings only of overseas born.

The leaves us with 19.5 million simulated Census records  $C$ ; the simulated nonresponse rate is more than 10% higher than that of the 2016 Census. Next, the additional non-respondents simulated are removed from the linked and unlinked census records  $C_L$  and  $C_{NL}$ , respectively. Finally, the FN linkage probabilities and the in-scope probabilities of the unlinked MADIP records are re-estimated given the adjusted Census-MADIP linkage outcome.

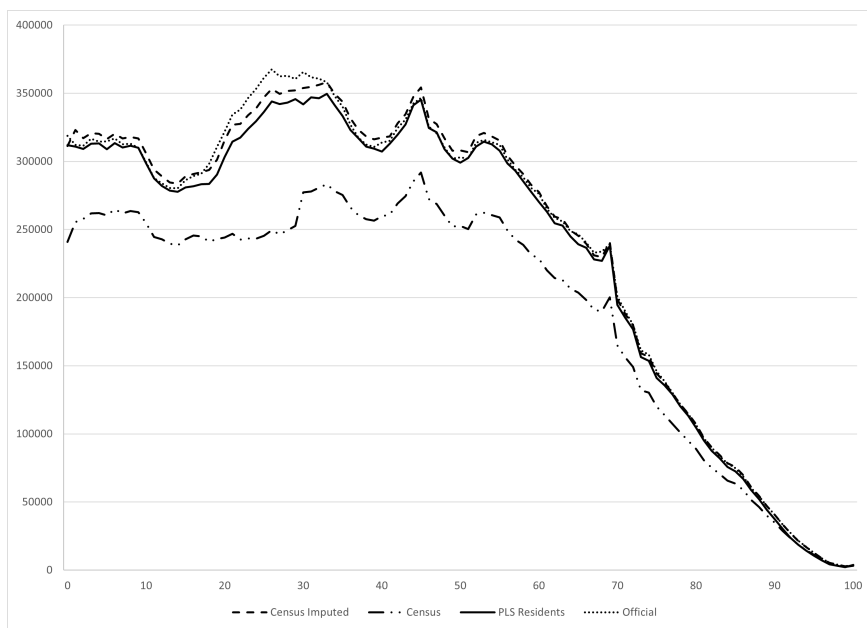


Figure 5: Thought experiment, counts by Census, Census-Imputed or PLS

Applying again the P14-method yields the imputed MADIP records  $U_{imp}$ . The corresponding PLS follows as  $C_L \cup U_{imp}$ , which contains 23.6 million records, and Census-Imputed follows as  $C \cup U_{imp}$  with 24.2 million records. Figure 5 plots the counts by age (in one-year groups). As expected, the Census counts are noticeably lower than in Figure 4, especially between the ages 15 and 29. It is interesting to notice that, without adjusting the imputation scheme given the additional Census nonresponse simulated, Census-Imputed and PLS counts still track the official counts well except for the 20-30 year olds.

Table 3: Enumeration and coverage errors (millions) in thought experiment. Census ( $C$ ), imputed MADIP records ( $U_{imp}$ ), Census-MADIP links ( $C_L$ ).

	$C$	$C \cup U_{imp}$	$C_L \cup U_{imp}$
Duplicated enumeration in $C$	0.3	0.3	N/A
+ Erroneous enumeration in $U_{imp}$	N/A	0.2	0.2
+ Double-counting by $C$ and $U_{imp}$	N/A	0.1	N/A
= Over-coverage (E)	0.3	0.6	0.2
Missing enumeration by $C$	5.0	N/A	N/A
Missing enumeration by $C_L \cup U_{imp}$	N/A	0.8	0.8
– TN of Census-MADIP linkage in $C \setminus C_L$	N/A	0.2	N/A
= Under-coverage (M)	5.0	0.6	0.8
Net coverage error (E – M)	-4.7	-0.1	-0.6
Gross coverage error (E + M)	5.2	1.2	1.0

Next, Table 3 gives the enumeration and coverage errors of Census, Census-Imputed and PLS in the thought experiment, similarly as Table 2 earlier. With or without the simulated additional 2.7 million non-respondents in Census, the gross coverage error is 2.5 or 5.2 million for Census, it is 0.8 or 1.0 million for PLS, and it is 1.0 or 1.2 million for Census-Imputed.

Clearly, imputing administrative records for the census non-respondents greatly reduces the enumeration errors than imputing the census respondent records. The resulting coverage errors of the imputed census databases (such as PLS or Census-Imputed) are robust against potentially a large increase of census nonresponse, since most of the census non-respondents can be found in the administrative sources by the appropriate scoping methods. Specifically in the Australian context, using MADIP records for Census imputation provides significant protection against potentially a low Census response.

## 4 Final remarks

We have developed methods for estimating the enumeration errors induced by imputing administrative records for census non-respondents in the presence of complications caused by linkage errors, and illustrated their uses in the context of the Australian Census 2016. This fills an apparent methodological gap, in order to meet the trend internationally for such uses of administrative data in population census. In particular, our approach to the estimation of FN linkage probability is more practical than the complicated and resource-demanding methods previously discussed in the literature of population size estimation based on capture-recapture data (e.g. Ding and Fienberg, 1994; Chipperfield and Chambers, 2015; Di Consiglio and Tuoto, 2018; Di Consiglio et al., 2019) which either requires a rematching study or extensive simulations of the linkage error mechanism. Moreover, the estimation of the proportion of in-scope administrative records is resilient against small or empty cell counts of the linked census records.

Insofar as some assumptions are needed for a statistical treatment of the

enumeration errors due to imputation, the quality of the relevant administrative data is important for successful applications. In Australia, the MADIP has a high coverage of people and dwellings, is timely, and has good coherence with Census and other national statistics (see Australian Bureau of Statistics 2023a, 2023b). In practical terms, this means that there is a relatively limited number of residents whose records are missing in  $A$ , although there may be many more records in  $A$  that do not belong to the target population  $U$ . The quality of administrative data sources concerns as well the features  $(z, y, x)$  required for the estimation models. Not all the countries may have equally good features, although the estimation method can be formally applied in the same manner. Attention to such quality aspects is important in practice.

An issue that is worth attention of further investigation and documentation concerns the enumeration errors at the lower levels of aggregation. Take e.g. the Australian Census 2016. The Australian Post Enumeration Survey (Chipperfield et al., 2017) provides an accurate count of residents only for broad areas such as state and age group. To assume that the corresponding enumeration errors are indicative of small area counts, a necessary condition is a high level of consistency for the small area geography classification between Census and MADIP. For example, suppose a person recently changed address and her true Census state of residence differs to that according to the outdated MADIP. If this person's record in  $U_{imp}$  is assigned to an incorrect small area because of this, it will be a missed enumeration in the correct area and an erroneous enumeration in the incorrectly assigned area.

We have considered small areas that typically contains between 3,000 and 25,000 residents each. Of the Census non-respondent records simulated in Section 3.5, about 88% had the same small area in  $C$  and  $A$ . This suggests that had the Census response rate been 80%, the effective response rate for the imputed Census at the small area level could have been about 96%, calculated as  $80\% + 20\% \cdot 90\% \cdot 88\%$  given 90% as the average proportion of in-scope records in  $U_{imp}$ , which is only a small decrease compared to 98% at the broad area level, i.e.  $80\% + 20\% \cdot 90\%$ . Thus, imputed Census small area counts can still achieve greatly reduced enumeration errors if the Census non-respondents are imputed with administrative records instead of the Census respondent records, despite the lack of definitive means to identify the in-scope records among those that are not linked to the Census respondents, or the fact that the small area geography is not always correct in the administrative sources.

However, more detailed investigation and documentation are beyond the scope of this paper, as well as extending the estimation methodology to lower levels of geography. This is naturally a topic for future research that will be useful to many statistical agencies given the importance of census population counts and the increasing uptake of administrative data in this context.

## References

- [1] Australian Bureau of Statistics (2016) Australian Demographic Statistics, Jun (2016) <https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/>

3101.0Explanatory%20Notes1Jun%202016?OpenDocument (accessed 13/12/2022)

- [2] Australian Bureau of Statistics (2020a), Multi-Agency Data Integration Project. <https://www.abs.gov.au/about/data-services/data-integration/integrated-data/multi-agency-data-integration-project-madip> (accessed 12/12/2022)
- [3] Australian Bureau of Statistics (2021a) Using administrative data to fill possible data gaps in the Census <https://www.abs.gov.au/statistics/research/using-administrative-data-fill-possible-data-gaps-census> (accessed 12/12/2022)
- [4] Australian Bureau of Statistics (2021b) Assessing administrative data quality to enhance the 2021 Census <https://www.abs.gov.au/statistics/research/assessing-administrative-data-quality-enhance-2021-census>
- [5] Australian Bureau of Statistics (2021c) 2021 Census overcount and undercount methodology <https://www.abs.gov.au/methodologies/2021-census-overcount-and-undercount-methodology/2021> (accessed 12/12/2022)
- [6] Australian Bureau of Statistics (2023a) Administrative data snapshot of housing, methodology. <https://www.abs.gov.au/methodologies/administrative-data-snapshot-population-and-housing-experimental-housing-data-methodology/30-june-2021> (accessed 12/11/2024)
- [7] Australian Bureau of Statistics (2023b) Administrative data snapshot of population, methodology. <https://www.abs.gov.au/methodologies/administrative-data-snapshot-population-and-housing-experimental-population-data-methodology/30-june-2021> (accessed 12/11/2024)
- [8] Brown J.J., Diamond I.D., Chambers R.L., Buckner L.J., Teague A.D. (1999). A methodological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society, Series A*, 162:247-267. doi:10.1111/1467-985x.00133
- [9] Chipperfield, J., Chu, R., Zhang, L.-C. and Baffour, B. (2024). Robust Statistical Estimation for Capture-Recapture using Administrative Data. *Journal of Official Statistics*, to appear.
- [10] Chipperfield, J., Brown, J. and Bell, P. (2017). Estimating the Count Error in the Australian Census. *Journal of Official Statistics*, 33:43-59. <http://dx.doi.org/10.1515/JOS-2017-0003>
- [11] Chipperfield, J. and Chambers, R. (2015). Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data. *Journal of Official Statistics*, 31:397-414. Doi:<http://dx.doi.org/10.1515/jos-2015-0024>

- [12] James Chipperfield, Noel Hansen, Peter Rossiter (2018). Estimating Precision and Recall for Deterministic and Probabilistic Record Linkage. *International Statistical Review*, 78:3-20. <https://doi.org/10.1111/insr.12246>
- [13] Di Consiglio L. and Tuoto, T. (2018). Population Size Estimation and Linkage Errors: the Multiple Lists Case. *Journal of Official Statistics*, 34:889-908. <http://dx.doi.org/10.2478/JOS-2018-0044>
- [14] Di Consiglio L., Tuoto, T. and Zhang, L.-C. (2019). Capture-recapture methods in the presence of linkage errors. In *Analysis of Integrated Data*, eds. L.-C. Zhang and R.L. Chambers. Chapter 3, pp. 39-72. Chapman & Hall/CRC.
- [15] Ding, Y. and Fienberg, S.E. (1994). Dual System Estimation of Census Undercount in the Presence of Matching Error. *Survey Methodology*, 20:149-158. <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1994002/article/14422-eng.pdf?st=YtHflfaV>
- [16] Dunne, J. and Zhang, L.-C. (2024). A system of population estimates compiled from administrative data only (with discussions). *Journal of the Royal Statistical Society, Series A*, 187:3-38. <https://doi.org/10.1093/jrssa/quad065>.
- [17] Farnell, J. and Darby, P. (2020). Administrative Data Informed Donor Imputation in the Australian Census of Population and Housing. *Statistical Journal of the IAOS*, 36:117-124.
- [18] Hogan, H. (1993) The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association*, 88:1047-1060.
- [19] Little, R.J.A. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.
- [20] Office for National Statistics (2001). *Census 2001: A Guide to the One Number Census*. ©CROWN COPYRIGHT. Available at <https://webarchive.nationalarchives.gov.uk/ukgwa/20160122034515/http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-conduct/the-one-number-census/methodology/step-by-step-guide/index.html>
- [21] Statistics New Zealand (2019a). Overview of statistical methods for adding administrative records to the 2018 Census dataset. <https://www.stats.govt.nz/methods/overview-of-statistical-methods-for-adding-admin-records-to-the-2018-census-dataset> (accessed 11/2021)
- [22] Statistics New Zealand (2019b). Data sources, editing, and imputation in the 2018 Census. <https://www.stats.govt.nz/methods/data-sources-editing-and-imputation-in-the-2018-census> (accessed 11/2021)
- [23] Statistics New Zealand (2019c). Dual system estimation combining census responses and an admin population. Available at <https://www.stats.govt.nz/methods/dual-system-estimation-combining-census-responses-and-an-admin-population/>

- [24] Statistics Canada (2021) Guide to the Census of Population, 2021, Appendix 1.7 – Use of administrative data to impute non-responding households in areas with low response <https://www12.statcan.gc.ca/census-recensement/2021/ref/98-304/2021001/app-ann1-7-eng.cfm> (accessed 7/2022)
- [25] Statistics Scotland (2022) Methodology Enhancements to Secure High Quality Census Outputs and Population Estimates. Available at <https://www.scotlandscensus.gov.uk/media/04sndppc/scotlandscensus-2022-methodology-enhancements-to-secure-high-quality-census-outputs-and-population-estimates.pdf> (accessed 12/2024)
- [26] Tietz, S., Mealar, A., Leather, F. and Dent, A. (2019) “Donor-based imputation methods for admin data: How to replace the number of rooms question on the Census”, *International Journal of Population Data Science*, 4(3). doi:10.23889/ijpds.v4i3.1299.
- [27] United Nations (2017a) Principles and Recommendations for Population and Housing Censuses Revision 3 [https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles\\_and\\_Recommendations/Population-and-Housing-Censuses/Series\\_M67rev3-E.pdf](https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf) (accessed 11/2021)
- [28] UNECE (2017b). Guidelines on the use of registers and administrative data for population and housing censuses, United Nations <https://unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20184.pdf> (accessed 11/2021)
- [29] United Nations (2021). Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses [https://unece.org/sites/default/files/2021-03/03\\_CensusAdminQuality\\_forConsultation\\_0.pdf](https://unece.org/sites/default/files/2021-03/03_CensusAdminQuality_forConsultation_0.pdf) (accessed 11/2021)
- [30] Wolter, K.M. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, 81:338-346. <http://dx.doi.org/10.1080/01621459.1986.10478277>
- [31] Zhang, L.-C. and Chambers, R.L. (2019). *Analysis of Integrated Data*. Chapman & Hall/CRC.