



## **Adjusting for linkage errors to analyse coverage of the Integrated Data Infrastructure (IDI) and the administrative population (IDI-ERP)**

Hochang Choi, Statistical Analyst, Stats NZ

Paper prepared for the 16<sup>th</sup> Conference of IAOS  
OECD Headquarters, Paris, France, 19-21 September 2018

Hochang Choi  
Hochang.choi@stats.govt.nz  
Stats NZ

**Adjusting for linkage errors to analyse coverage of the Integrated Data Infrastructure (IDI) and the administrative population (IDI-ERP)**

DRAFT VERSION 10/09/2018

Prepared for the 16<sup>th</sup> Conference of the  
International Association of Official Statisticians (IAOS)  
OECD Headquarters, Paris, France, 19-21 September 2018

**Note:**

This Working Paper should not be reported as representing the views of the Statistics New Zealand. The views expressed are those of the author.

## ABSTRACT

The Integrated Data Infrastructure (IDI) is a linked data environment combining a variety of administrative and survey datasets. It is anchored by a spine which defines an “ever-resident” population constructed from the union of births, tax and visa data. Stats NZ has constructed an experimental New Zealand resident population (the IDI-ERP) from the IDI by defining rules to identify people who are likely to be usual residents and exclude people likely to no longer be usual residents in New Zealand at some reference date. The IDI-ERP may be of interest as a reference population for small area population estimation, and other social and population research. It is therefore of interest to analyse the coverage, or representativeness, of the IDI-ERP with respect to the true usual resident population of New Zealand. Linkage of the 2013 Census data to the IDI spine provides an opportunity for detailed analysis of the population coverage patterns of the IDI-ERP. While straightforward in principle, this analysis is complicated by error in the linkage of the census to the IDI spine and the fact the census itself is subject to under-coverage with respect to the true usual resident population. In this paper, we outline our approach for adjusting IDI-ERP estimates for these sources of error and present preliminary estimates of the population coverage of the IDI-ERP, by sex, age and ethnicities.

Keywords: Administrative data, Bayesian Inference, Population estimation, Missing data, Linkage error

## 1. Introduction

This paper looks at a method for adjusting for linkage errors to analyse coverage of the Integrated Data Infrastructure (IDI) and the administrative population (IDI-ERP). Section 1 introduces the IDI and IDI-ERP, while also explaining false negative and false positive links. We present how different national statistics offices are managing linkage errors in section 2. Section 3 demonstrates a method of adjusting for linkage errors. We look at the preliminary estimates of the coverage of the IDI and IDI-ERP, by sex, age, and ethnic groups in section 4. We conclude in section 5 with some areas of future work.

### 1.1. Background: Stats NZ's Integrated Data Infrastructure (IDI) and administrative population (IDI-ERP)

Stats NZ is working actively towards a future census based primarily on Government's administrative data, supported by redevelopment of its household surveys.

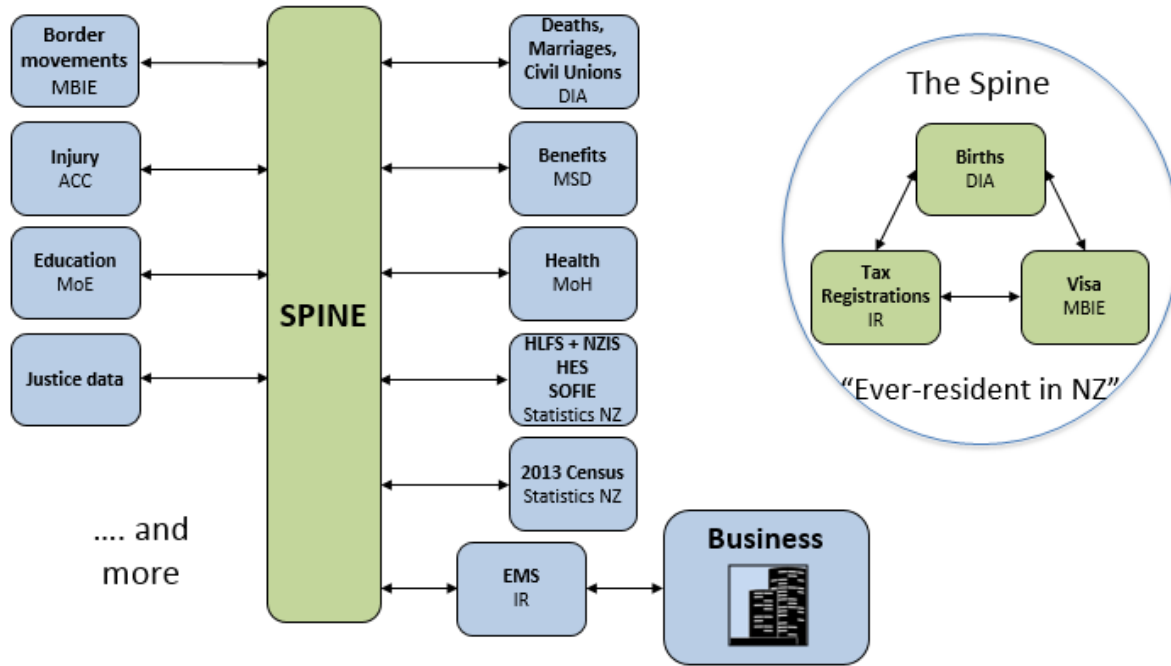
The Integrated Data Infrastructure (IDI) is a linked data environment combining a variety of administrative and survey datasets. It allows for statistical outputs and research on the transition and outcomes of people through education, health and safety, and migration data. It is anchored by a spine, defining an "ever-resident" population of New Zealand. The spine is constructed from the union of births, tax and visa data (see Fig. 1).

Stats NZ has constructed an experimental New Zealand population (the IDI-ERP) from the IDI. We identify all individuals with activity in administrative data source such as tax, health and education two years prior to a chosen reference date. We exclude individuals who deceased and those who migrated overseas before the reference date.

The primary motivation for constructing this population estimation was to inform research at Stats NZ on the prospects for using administrative data in place of census data to underpin small area population estimation. The IDI-ERP may also be of interest as reference population for other social and population research. It is therefore of interest to analyse the coverage, or representativeness, of the IDI-ERP with respect to the true usual resident population of New Zealand.

Linkage of the 2013 Census data to the IDI spine provides an opportunity for detailed analysis of the population coverage patterns of the IDI-ERP. The difference between the IDI-ERP and the true population is called a coverage error. Some of these errors are due to genuine under-/over-coverage in the IDI-ERP. We define as genuine under-coverage in the IDI-ERP if an individual in the census does not appear in the IDI-ERP. For example, if individuals in the census do not interact with government agencies, then they will not appear in the IDI-ERP. We define as genuine over-coverage in the IDI-ERP if an individual in the IDI-ERP should not respond to the census. For example, if individuals are away from New Zealand at census night and did not complete a census form but interact with government agencies by enrolling in education, then they should not appear in the census.

**Figure 1 The Integrated Data Infrastructure (IDI)**



## 1.2. False positive and false negative links

While identifying genuine coverage error is straightforward in principle (Graham, & Lin, 2017), the analysis is complicated by error in the linkage of census to the IDI spine. Stats NZ (2006) use name, sex and birthday for data integration at unit record level where it introduces linkage errors. This paper focus on the linkage errors between the census and the spine, and two types of linkage errors, false negatives and false positives.

We define false negatives as census records that should have been linked to the spine record, but were not. For example, a person may use a nickname in the census but use a legal name in the administrative data and hence may not be linked. The second type of linkage error is false positives where the census record have been incorrectly linked to the spine record. For example, John Doe in the census is linked to John Doo in the spine when they are in fact two different people.

## 2. Linkage error in different national statistics offices

The combined use of administrative data from different sources is an opportunity for researchers and national statistics offices to exploit more. However, linkage errors exist in data integration at micro level. Many national statistics offices evaluate information regarding the matching data to provide accuracy and credibility of the data.

### 2.1. Italian National Statistical Institute

Tuoto (2016) from Italian National Statistical Institute proposed a new method for linkage error estimation by enhancing the Fellegi and Sunter (1969) procedure to evaluate the linkage errors. Effectiveness of the proposed method is evaluated using the synthetic data where the true matching status is known. The data has been created to mimic reality – presence of linkage errors.

They first select a training sample (10% of the original data) and run model selection procedures. Then the best model for predicting the matching status and linkage errors is identified, and it is applied on the full data. Since the true matching status is known, the estimators are analysed and compared. These estimators provide a quality indicator of the linked data.

## **2.2. Australian Bureau of Statistics (ABS)**

ABS (*Kindermann et al 2016*) proposed a model-based method to estimate the precision for record linkage. The precision is a true positive rate, and is a quality indicator of the linked data which is useful in the planning and analysis of record linkage process.

There are two approaches on measuring precision, and the first one is to simulate the linking processing many times. Then they look at the agreement pattern between data sets during the simulation based on underlying probabilities. The second approach is to obtain an algebraic estimator. Their result suggests that the estimators of precision perform well in both the synthetic and real data. This would provide efficient and cheaper alternatives without relying on traditional clerical review.

## **3. Methods**

Stats NZ has performed many data integration projects and one of the projects was related to the linkage between the census and the spine. The current link rate between the census and the spine is 94% (so 6% of census is not linked to the spine). This seems surprisingly low as the spine aims to capture most of ever-residents. We understand this is due to false negative links. We have discovered that the group who filled the census in paper form had a census to the spine link rate of 91.5%. This was lower than the link rate for those who filled the census electronically (98.4%). The paper filled census records had poorer quality in names and addresses as they were scanned which allowed less links to be made.

In this section, we present a method to adjust for false negative links. The steps of the method include imputing missing values in section 3.1, correcting the IDI spine indicator in section 3.2, correcting the IDI-ERP indicator in section 3.3 and accounting for over-coverage in the IDI-ERP in section 3.4.

### **3.1. Imputing missing values**

The census dataset includes missing values in country of birth, region, studying status, labour force status, and highest qualification. We need to impute values for these variables where they are missing since they are used in the linkage error correction model. For each variable with missingness, we model the probability of being missing from the fully observed variables. We then use the probabilities to identify the nearest neighbours for each missing values and use a Bayesian Bootstrap (*Rubin, 1981*) procedure to sample an imputed value among the nearest neighbours.

### **3.2. Correcting the IDI spine indicator**

The IDI spine indicator flags if a census record is linked to the spine. We applied a method to generate a corrected spine indicator after adjusting for the effect of linkage error. In reality, clerical review identified false positive links between the census and the IDI spine with the estimated rate of 1.54% and sampling error of 0.32%. However, in this analysis, our assumption is that there are no false positive links – meaning that all links between the census and spine are true matches. Thus we are only correcting for false negatives – individuals in the census that are not linked to the spine, but ought to be.

A false negative rate in the census-spine link is not easily obtainable as it is difficult to identify true matches that are not linked. We use the following seven steps to produce a false negative indicator for individuals in the census who are not matched to the spine:

1. The census is divided into two groups based on their census responses,  $M$  and  $M'$ , where  $M$  is a group that should be in the spine, and  $M'$  is a complement of the group  $M$ . The following criteria is designed to obtain  $M$  from the census:
  - New Zealand born individual with taxable income on census night – wages, salaries, New Zealand superannuation, veterans pensions, sickness benefits, domestic purposes benefits, unemployment benefits, and student allowances,
  - Individual arrived in New Zealand after July 1997 with taxable income on census night.

The spine is constructed from the union of three sources, births, tax and visa data. Our aim is to capture individuals in the census who are very likely to be in the spine by identifying those in two spine sources (a stricter spine).

2. For each individual in  $M$ , we assign the false negative indicator a value of 1 if an individual is not linked to the spine, and 0 otherwise.
3. Estimate the false negative ( $FN$ ) probabilities,  $Pr(FN | M, X)$  for individuals in  $M$  – the probability an individual in  $M$  was not linked to the spine – as a function of covariates  $X$  using a logistic model. Since we are assuming that all individuals in  $M$  are truly in the IDI spine, all non-links in this group must be false negatives. Conditional on the covariates, we assume the false negative probability is the same for the  $M$  and  $M'$  groups. Unfortunately there is no way of testing this assumption.
4. Estimate the probability of non-linkage ( $N$ ) for the  $M'$  group,  $Pr(N | M', X)$  as function of covariates using a logistic model.
5. For unlinked individuals in the  $M'$ , we obtain the probability that they are false negative links from the false negative probabilities obtained in step (3) and the negative probabilities for  $M'$  obtained in step (4), using

$$\begin{aligned} Pr(FN|N, M', X) &= \frac{Pr(FN, N|M', X)}{Pr(N|M', X)} \\ &= \frac{Pr(FN|M', X)}{Pr(N|M', X)} \\ &= \frac{Pr(FN|M, X)}{Pr(N|M', X)} \end{aligned}$$

Where the last equality follows from the assumption that the false negative probabilities estimated for the  $M$  group holds also for the  $M'$  group.

6. Generate false negative indicators for unlinked  $M'$  records using the probabilities obtained in step (5).
7. Combine false negative indicators for individuals in both  $M$  and  $M'$  from step (2) and step (6).

We use these false negative indicators to produce a corrected IDI spine indicator. Any individual with a false negative indicator equal to 1 will also be assigned a corrected IDI spine indicator of 1. Our assumption in this method is that we treat individuals in  $M$  who are not linked to the spine as false negative links.

### 3.3. Correcting the IDI-ERP indicator

The IDI-ERP indicator flags if a census record is included in the IDI-ERP. Some census records genuinely do not appear in the IDI-ERP due to the current activity rules. Hence we need to account for genuine under-coverage when we correct the IDI-ERP indicator.

We have already identified false negative links while correcting the spine indicator. These census records do not have the associated spine records as they are not in the spine, but ought to be. Thus there is no information on whether these census records should be included in the IDI-ERP or not. We use the following method to correct the IDI-ERP indicator:

1. Identify a true match group, that is, individuals in the census with a spine indicator and corrected spine indicator equal to 1.
2. Calculate the probabilities of being in the IDI-ERP for individuals in the true match group. This provides a measure of genuine under-coverage, and we are assuming the coverage patterns for the false negative groups are the same as for individuals in the true match group.
3. Use logistic regression to generate IDI-ERP indicators for false negative group using probabilities from step (2).

### 3.4. Accounting for over-coverage in the IDI-ERP

Once the IDI-ERP indicators for the false negative group in the census are corrected, we then need to find a matching record from the spine which were not linked to the census. The matching record should have the same characteristics – age, sex, region, and IDI-ERP indicator. This matching exercise allows us to account for the over-coverage in the IDI-ERP – individuals included in the IDI-ERP but not the census.

**Table 1 Cross tabulation of the census and the IDI-ERP before the matching**

		IDI-ERP before the matching	
		1	0
Census	1	$N_{11} - \alpha$	$N_{10} + \alpha$ (false negatives)
	0	$N_{01} + \alpha$ (over-coverage in the IDI-ERP due to linkage errors)	

**Table 2 Cross tabulation of the census and the IDI-ERP after the matching**

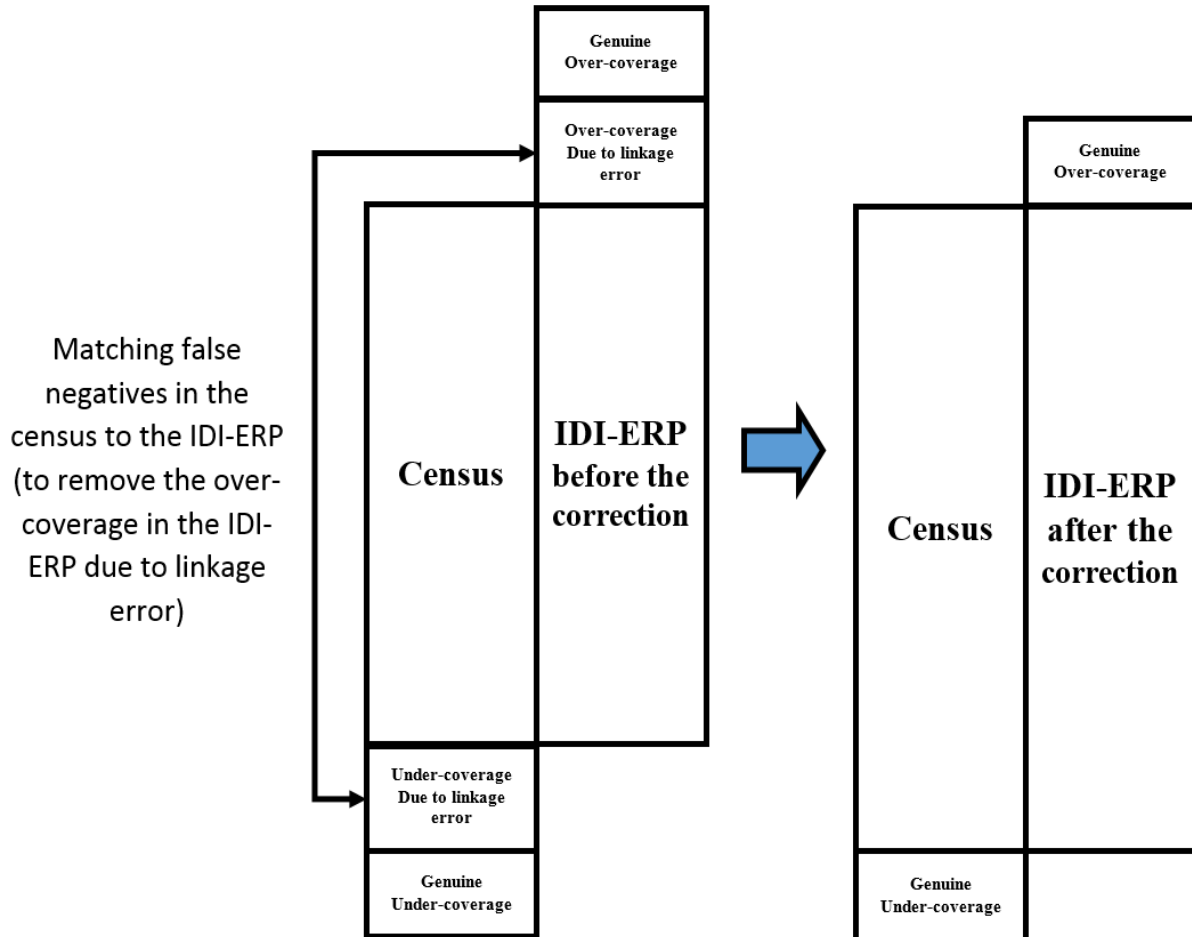
		IDI-ERP after the matching	
		1	0
Census	1	$N_{11}$	$N_{10}$
	0	$N_{01}$	

Tables 1 and 2 show the cross tabulation of the census and the IDI-ERP before and after the matching, respectively. By applying the IDI-ERP corrections in section 3.3., we can identify false negatives ( $\alpha$ ) in the census. These census records will move from the (1, 0)-cell to the (1, 1)-cell. On its own, this correction would lead to over-stating the over-coverage in the IDI-ERP since associated  $\alpha$  still exist in (0, 1)-cell.



Therefore, associated IDI-ERP records from the (0, 1)-cell would also need to move to the (1, 1)-cell (see Fig. 2).

**Figure 2 Union of the census and the IDI-ERP before and after the matching**



## 4. Results

### 4.1. Coverage of the spine and IDI-ERP

Figure 3 shows the percentage of individuals in the census who were found in the spine, before and after adjusting for linkage error. The coverage rate increased after accounting for linkage error by at least 2% for all age groups, with a maximum increase of 6% for age 18 to 25 which had poor coverage rates before the correction. Hence this could mean that false negatives in the census-spine link are more likely to occur for ages 18-25. So that the perceived low coverage of the spine for this age group, is, in part due to linkage error, and under-estimates the true coverage of the spine.

**Figure 3 Percentage of individuals in the census who were linked to the spine, before and after correcting for the linkage error**

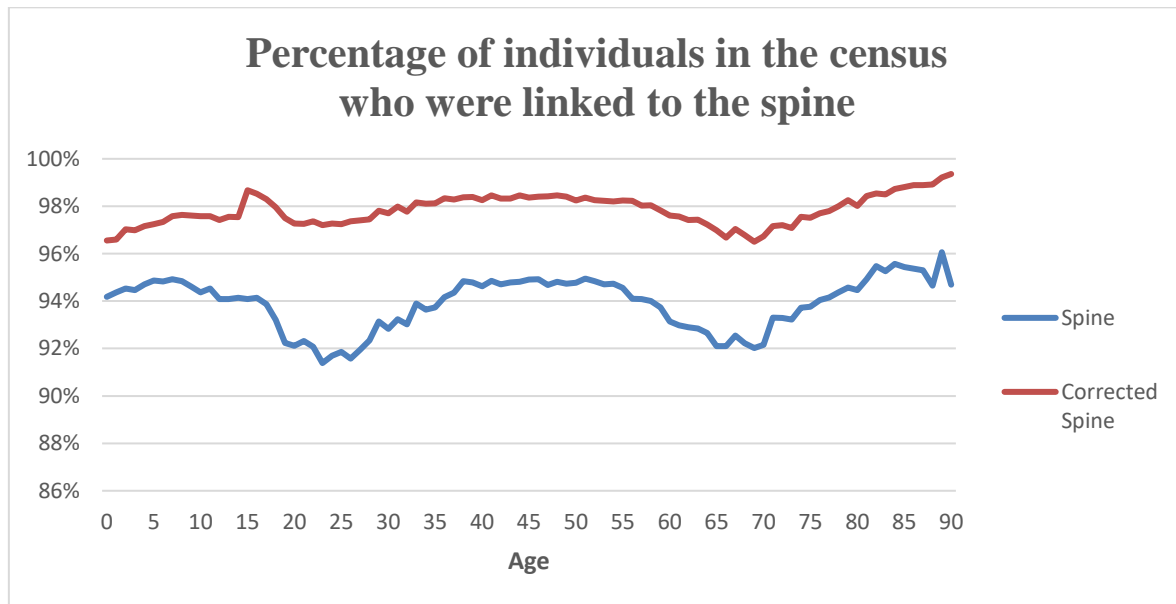


Figure 4 shows the percentage of individuals in the census who were also included in the IDI-ERP, before and after correcting for linkage error. The improvement in the coverage rate for the IDI-ERP is not as significant as the spine, but the coverage rate is above 90% for all age groups after the correction. The coverage of individuals aged 5-15 years is close to 98% after the correction. This corresponds to many children attending school, and thus they should be captured in the IDI-ERP. We observe a noticeable decrease for young adults with the coverage dipping below 92% after the correction for ages 18–25. This can be due to some people not interacting with government agencies.

**Figure 4 Percentage of individuals in the census who were linked to the IDI-ERP**

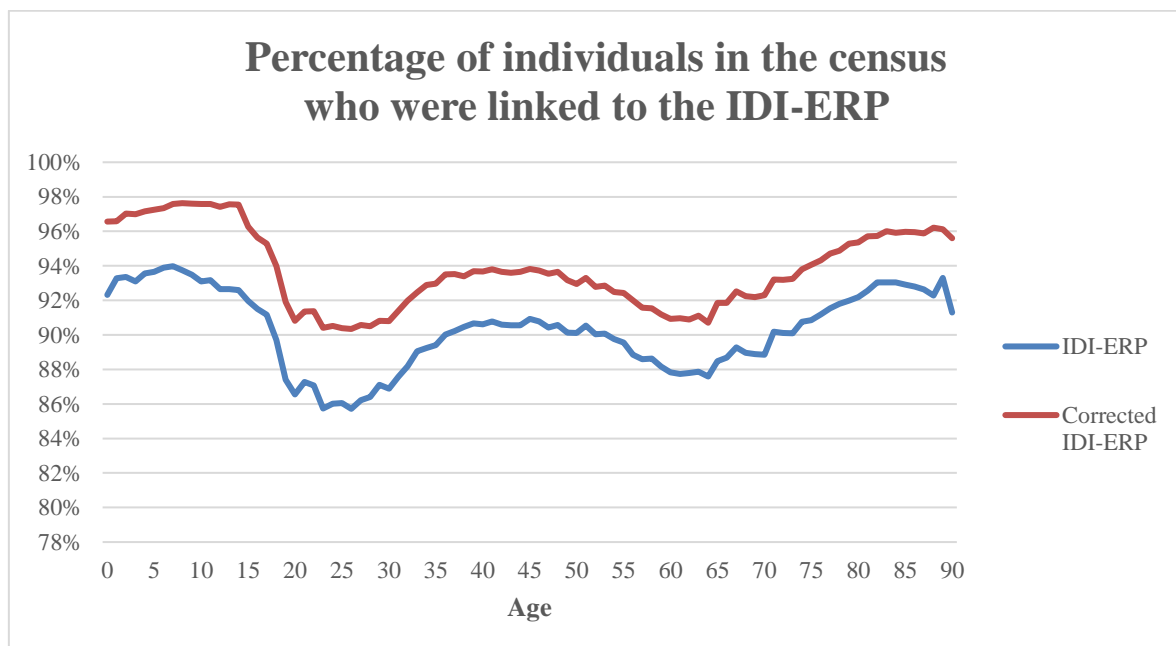


Figure 5 shows the percentage difference between the IDI-ERP and the corrected IDI-ERP, by sex and age. This result again suggests false negatives in the census-spine link are more likely to occur for younger age groups (aged 15-30). Both male and female has similar patterns where older age groups (aged 60-80) are more prone to false negatives. Quality in the census and administrative data, such as names, may have hindered these age groups to successfully match the census to the spine.

**Figure 5 Percentage difference between the IDI-ERP and the corrected IDI-ERP, by sex and age**

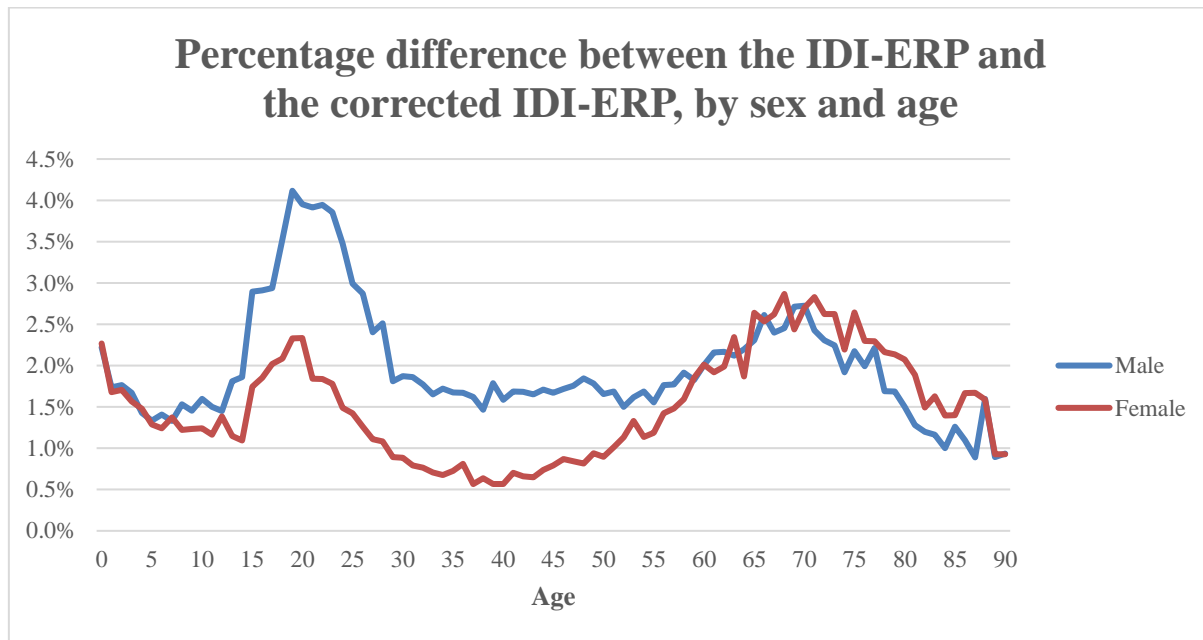
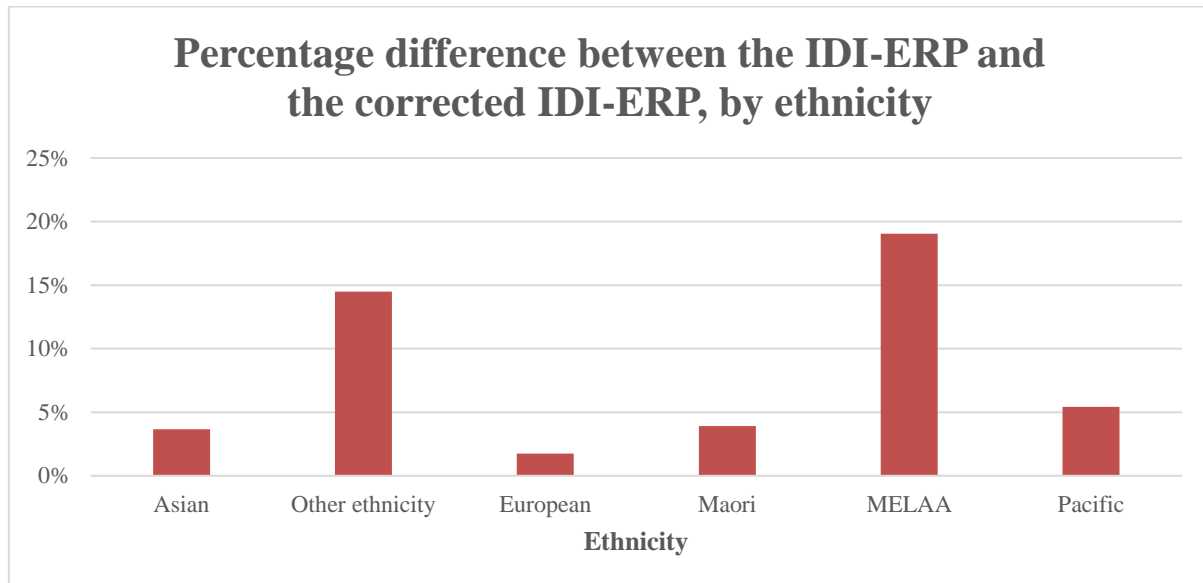


Figure 6 shows the percentage difference between the IDI-ERP and the corrected IDI-ERP, by ethnicity. The result shows that there is about 15% difference between the IDI-ERP and the corrected IDI-ERP for other ethnicity. This ethnic group includes a residual category such as non-response and we assume the census data quality for this group may not be a good enough to link to the spine. The highest percentage difference of 20% occurred for Middle Eastern, Latin American or African group. The magnitude of the difference is surprising, however, this group has a relatively small count so that any change in count due to linkage error corrections would contribute a large change in the percentage difference.

The difference between Europeans and the other ethnic groups are evident. We found that proportionally more Europeans have filled in electronic census form than the other ethnicities. The data quality would be better and more links were made; hence less adjustment for linkage errors.

**Figure 6 Percentage difference between the IDI-ERP and the corrected IDI-ERP, by ethnicity**



## 5. Conclusion

Linkage errors are often considered as negligible and thus ignored. However, we have shown in this paper that the linkage errors must be addressed. We provided a method to identify any individuals in the census that should have been linked to the IDI-ERP but were not. We then obtained the IDI-ERP indicators for census records and adjusted for over-coverage in the IDI-ERP due to linkage errors. We produced initial results to show the impact of the method.

Our results indicate that particular sub-groups (for example, by age, and ethnicity) in the census may be prone to non-match to the spine due to their characteristics. This shows that the analysis on coverage of administrative population may be biased if we do not adjust for the linkage error.

There are several limitations to our methodology. We assumed that there are no false positives – individuals in the census who are falsely linked to the IDI-ERP – which clerical review found that this assumption is not valid. We also assumed that all individuals in the census M group are truly in the IDI spine. There was a computational limitation which did not allow us to use multiple imputation method to account for uncertainties for running the model.

### 5.1. Recommendations for future work

We provide recommendations for future work which would improve the results in this paper. Model can be improved by accommodating false positives in the correction model. We can extend our methodology by using a multiple imputation method to account for uncertainties in the IDI-ERP estimates when adjusting for linkage errors.

Our methodology relies on few assumptions. Thus it is important to assess the methodology by running a sensitivity analysis to check the impact of our assumptions on our results. It is also important to compare the results we obtained from the model with the true matching status. Practically it is difficult to know the true matching status. However, we can run an extensive clerical review to get a good estimates of the false negative rates and false positive rates, and the estimates can be treated as a true matching status. Lastly, we need to further investigate on sub-groups, such as by age, sex, ethnicities, to identify the underlying

linkage error pattern. The pattern may be useful for improving the data quality which would reduce the error when linking.

## **6. References**

Graham, Patrick, & Lin, Anna, 2017, Small domain population estimation based on an administrative list subject to under and over-coverage. ISI2017 Marrakech.

Tuoto, Tiziana, 2016, New proposal for linkage error estimation. Statistical Journal of the IAOS 32, p 413-420.

Kindermann, Bindi, & Chipperfield, James, & Hansen, Noel, & Rossiter, Peter, & Wright, Jeffrey, 2016, Measuring precision for deterministic and probabilistic record linkage. IJPDS 2017 Issue 1, Vol 1:091.

Fellegi, Ivan, & Sunter, Alan, 1969. A theory of record linkage. Journal of the American Statistical Association 64, p 1183–1210.

Rubin, Donald, 1981. The Bayesian Bootstrap. The Annals of Statistics, Vol. 9, No. 1, p 130-134.

Stats NZ, 2006, Data Integration Manual. Available from [www.stats.govt.nz](http://www.stats.govt.nz).