Australian Bureau of Statistics

1351.0.55.043

Research Paper

Assessing the Quality of Linking Migrant Settlement Records to 2011 Census Data

New Issue

Research Paper

Assessing the Quality of Linking Migrant Settlement Records to 2011 Census Data

Kate Richter, Gokay Saher and Paul Campbell

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 19 AUG 2013

ABS Catalogue no. 1351.0.55.043

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper. For further information, please contact Dr Phillip Gould, Analytical Services Branch, on Canberra (02) 6252 5315 or email <analytical.services@abs.gov.au>.

CONTENTS

ABBREVIATIONS

ABS	Australian Bureau of Statistics
ASB	Analytical Services Branch
CDE	Census Data Enhancement
DIAC	Department of Immigration and Citizenship
RSE	Relative Standard Error
SACC	Standard Australian Classification of Countries
SDB	Settlement Database
SDB–Census	Refers to the dataset formed by linking the SDB to the Census
SLCD	Statistical Longitudinal Census Dataset

ASSESSING THE QUALITY OF LINKING MIGRANT SETTLEMENT RECORDS TO 2011 CENSUS DATA

Kate Richter, Gokay Saher and Paul Campbell
Analytical Services Branch

ABSTRACT

In 2006, as part of the Australian Bureau of Statistics' Census Data Enhancement (CDE) project, the ABS conducted a Migrants Quality Study. The purpose of this study was to assess the feasibility of linking the Department of Immigration and Citizenship's Migrant Settlement Database (SDB) to the Statistical Longitudinal Census Dataset (SLCD), a 5% sample of the 2006 Census of Population and Housing, without the use of name and address as linking variables. The details of the process are set out in the ABS research paper, *Assessing the Quality of Linking Migrant Settlement Records to Census Data* (Wright, Bishop and Ayre, 2009).

The 2006 study linked the SDB to the 2006 Census and analysed the subsequent file. Analysis indicated that linking the SDB to the SLCD was feasible and could produce useful information that no other data source could provide. Issues were identified pertaining to the quality of linked data, and a recommendation made that they be thoroughly explained to ensure that the linked dataset was correctly understood, interpreted and appropriately used.

The 2011 Migrants Quality Study was conducted with the aim of creating and evaluating a linked SDB–2011 Census dataset, with a focus on assessing the modifications and improvements to the linking process arising from the 2006 Study.

This paper provides some background to the 2006 and 2011 Migrants Quality Studies, a brief description of the linking strategy and process, a discussion of the quality of the linking of the SDB to the 2011 Census, and introduces its potential as an analysis dataset.

1. INTRODUCTION

The Census Data Enhancement (CDE) project is aimed at integrating unit record data from the Australian Bureau of Statistics' (ABS) Census of Population and Housing with other ABS and non-ABS datasets to create new datasets for statistical and research purposes. Several CDE projects were undertaken using the 2006 Census and are now being continued using the 2011 Census.

Initially the CDE projects assessed the feasibility of bringing together the full Census dataset with other datasets without name or address, in line with the ABS's original proposal and the *Australian Statistician's Statement of Intention* (ABS, 2005). The establishment of the CDE project and associated details are presented in ABS information papers *Census Data Enhancement Project: An Update* (ABS, 2006b) and *Enhancing the Population Census: Developing a Longitudinal View* (ABS, 2006a).

An update on the outcomes of the 2006 CDE project and plans for the continuation of the project for 2011 were presented in the ABS information paper, *Census Data Enhancement Project: An Update, October 2010* (ABS, 2010a).

CDE projects at the ABS aim to generate two key products:

- A Quality Study, providing an assessment of the feasibility of the linking methodology, data quality and the data's fitness for purpose; and
- A Statistical Study, linking two datasets with the aim of publishing results and providing a range of enhanced Census data for use by the wider community.

In line with the original proposal and Statement of Intention, the ABS undertook several Quality Studies. The Quality Studies aimed to investigate the feasibility and likely quality of linking datasets when name and address are not available. This approach involves linking data using two information sets and comparing the resulting datasets. First, datasets are linked using name and address in addition to the array of other available linking information. The resulting dataset is termed a *Gold Standard* linked dataset. The same datasets are then linked without the use of name and address in any phase of the linking, and the resulting dataset is termed a *Bronze Standard* linked dataset. The current Statement of Intention specifies that the Gold Standard linked datasets will be deleted following an evaluation period.

Consequently, during this period the Bronze Standard is compared against the Gold equivalent in order to ascertain whether linking the datasets without name and address can produce a dataset of sufficient quality to be used confidently in analysis.

The 2006 Migrants Quality Study assessed the feasibility of probabilistically linking the administrative data from the Department of Immigration and Citizenship's (DIAC) Settlement Database (SDB) to the 2006 Statistical Longitudinal Census Dataset (SLCD) without the use of name and address. SDB person records were probabilistically linked to the 2006 Census using both Gold and Bronze linking strategies. The Quality Study determined that probabilistic linking was feasible, however a number of quality issues with the SDB data were identified in this evaluation (Wright, Bishop and Ayre, 2009). These issues, described in detail in Section 2 below, led directly to a number of initiatives aimed at improving the quality of the SDB dataset prior to the 2011 Migrants Quality and Statistical Studies. An additional key aim of the 2011 Migrants Quality Study is to assess the impact of these initiatives.

The SDB–Census linking project aims to append some key variables of interest from the SDB to the Census for the migrant population. These variables include visa class (skilled, humanitarian and family visas), application status (primary or secondary applicant) and whether the application was processed onshore or offshore. The resulting linked dataset can be used to assess a range of questions, notably the relationship between a migrant's visa class and their post-arrival social and economic outcomes.

It is envisaged that integrated Census and SDB data will serve to provide detailed information on recent permanent settlers in Australia in relation to variation in family formation, labour market, housing and other socioeconomic outcomes across different migrant sub-groups. In the longer term, this rich data source could assist in the future development and evaluation of immigration programs and support services for migrants. The potential for longitudinal analysis using the SLCD linked to the SDB could lead to improved research into, and identification of, the causal factors underlying particular migrant outcomes.

This paper provides a summary of the 2011 Migrants Quality Study. It first briefly describes and evaluates the 2006 study, which provided a basis for the 2011 study. It then introduces the datasets to be linked, before outlining both the probabilistic linking methodology generally and the methodology applied in the 2011 Migrants Quality Study. Particular attention is paid to the improvements made as a direct result of the evaluations of the linkage in the 2006 Quality Study. The paper finally presents an evaluation of the linked datasets, comparing the Gold Standard and Bronze Standard linked datasets. Readers interested only in the assessment of the quality of the 2011 SDB–Census linked dataset should proceed directly to Section 5.

2. THE 2006 MIGRANTS QUALITY STUDY

The 2006 Migrants Quality Study aimed to investigate the feasibility of linking Department of Immigration and Citizenship's (DIAC) Migrant Settlements Database (SDB) with the Census, and accordingly with the five Statistical Longitudinal Census Dataset (SLCD)), a 5% sample of the 2006 Census of Population and Housing. The 2006 Census processing period provided an opportunity to link the SDB to the full Census dataset using two approaches, first utilising and then omitting name and address as linking variables, creating a Gold and Bronze Standard linked dataset respectively. Linking with name and address, while not perfect, provides a valuable benchmark for assessing linkage quality when name and address are not available. Results were mixed but generally indicated that a linking project involving SDB and Census data could be further pursued. A description and detailed evaluation of the study can be found in Wright, Bishop and Ayre (2009). This section provides a cursory evaluation, focusing on the points which directly impact the 2011 Quality and Statistical Studies.

The 2006 SDB consisted of 806,952 records, comprised of migrants who obtained an Australian visa between 1 January 2000 and Census night, 8 August 2006. The 2006 Migrants Quality Study linked this file to the 2006 Census (19,050,146 records). Linking using name and address resulted in 63% of the SDB records being linked to a Census record with high accuracy. Linking without name and address linked 66% of SDB records, of which 80.2% were identical to a Gold Standard link.

For both the Gold and Bronze datasets, the proportion of records linked was lower than that in other Census Data Enhancement quality studies, such as the Indigenous Mortality Quality Study and Simulated SLCD Study. However, both Gold and Bronze datasets were found to be broadly representative of the SDB, making the files potentially useful for analysis.

The ABS Analytical Services Branch undertook an in-depth analysis and identified a number of causes for the less than optimal linking outcome. First, although the SDB is theoretically a subset of the Census, a significant number of recent migrants were overseas on Census night, and hence could not be linked. Some further recent migrants did not appear on the Census due to Census undercount. Original estimates of mortality, undercount, and both permanent and temporary emigration suggested that of the 806,952 SDB records, around 107,000 were out of scope on the 2006 Census (Wright, Bishop and Ayre, 2009, p. 12). However, newly available data on the 2011 SDB was able to accurately identify individuals overseas at a particular point in time. This data applied to the 2006 SDB, indicates that up to 170,000 recent migrants were out of scope in the 2006 Migrants Quality Study, suggesting the project was of a higher quality than originally reported.

Second, addresses on the SDB were often out of date or incomplete. Incorrect address details are frequently due to the subsequent internal migration of recent migrants after arrival and the lack of avenues for obtaining a more recent address. Analysis suggested approximately 100,000 SDB records could not be linked due to incomplete addresses. Similarly, a number of given names and family names were poorly reported. In some instances, names were misplaced, that is first names were entered in the family name field and vice versa, hindering comparison on name fields. In some further cases the first name was missing, but potentially both first and family name had been entered in the family name field. There were 7,171 such records, primarily from India, Indonesia and Afghanistan. Finally, one third of unlinked records were missing four or more linking variables and one quarter had five or more linking variables missing (Wright, Bishop and Ayre, 2009, p. 14).

These issues led to the development of a number of new methods aimed to improve both the quality of SBD–Census linking, and the accuracy of evaluating the resulting linked datasets. These methods are discussed in Section 3 and Section 4.2.

Finally, it is worth noting that, although the 2006 Migrants Quality Study was aimed solely at assessing the feasibility of the project, the National Migrant Statistics Unit (NMSU) released nine data cubes and two articles in 2010 presenting data and analysis highlighting the potential of the 2006 Statistical Study, *Settlement Outcomes for Humanitarian Program Migrants – Experimental Estimates from the Migrants Statistical Study* (ABS, 2010b) and *Economic Outcomes of Skilled Program Migrants – Experimental Estimates from the Migrants Statistical Study* (ABS, 2010c).

3. THE 2011 DATASETS

This section provides an overview of the two datasets being linked, namely the Migrant Settlement Database (SDB) and 2011 Census.

3.1 Migrant Settlement Database

The SDB is compiled by the Department of Immigration and Citizenship (DIAC) from a number of data sources. The SDB extract used in the 2011 Migrants Quality Study covered the period 1 January 2000 to 9 August 2011 (Census night) and contained the records of 1,649,260 persons who, during that period, were granted visas to live permanently in Australia. Persons on the SDB comprise skilled, humanitarian (including refugee), and family migrants. The SDB excludes temporary visa holders and non-visa settlers, such as persons from New Zealand.

The SDB dataset comprises both onshore and offshore applicants. A person who applies offshore and is granted a permanent entry visa will be given a grant (approval) number. The person is also given a visa evidence number when they present their passport to be stamped with the necessary visa to enable them to travel into Australia. The person's grant (approval) number, date of grant, and visa evidence number are added to their records on the SDB. They do not have an SDB arrival date until they arrive in Australia. If the person does not arrive in Australia within 13 months, their records will be flagged as a 'non-arrival'. Arrival date was therefore used to remove persons from the SDB who have been granted permanent entry visas but have not yet arrived in Australia.

For a person who applies onshore for a permanent resident visa to remain in Australia, the arrival date listed on the SDB is the date of their last entry into Australia. They have an SDB approval number and date of approval, but do not necessarily have a visa evidence number unless they present their passport to be stamped.

DIAC provided two versions of the SDB for the 2011 Quality Study, the first in December 2011 and the second in August 2012. The first version was used to begin developing a linking strategy, whereas the second contained a number of improvements and auxiliary variables. First, DIAC provided flag variables from their Travel and Immigration Processing System (TRIPS) data, indicating whether, and if so when, migrants had last departed the country. The movement date and direction were used to identify and remove 315,998 (19.2%) records from the original file of 1,649,260 migrants arriving between 2000 and 2011. Second, DIAC utilised Medicare data to update address fields. Third, TRIPS data were used to evaluate and repair name fields. Additionally, the SDB itself was used to create an index of common migrant given and family names. This index was used to help repair records where full name was reported in the family name field.

Further analysis of the SDB extract identified and removed 17,655 Australian born persons, 401 deceased persons and 158 duplicate records. The Australian born persons may be individuals who, in virtue of staying overseas for too long a period, lost their permanent residency. Duplicate records presumably represent persons who made multiple visa applications and who received multiple acceptances, or possibly offshore applicants who applied for and changed their visa status after arrival in Australia. The deceased persons identified here are only a small proportion of those estimated to have died between arrival and the 2011 Census. The resultant file for linking contained 1,315,048 records.

3.2 Census

The 2011 Census file used for this study consisted of 20,928,304 records, excluding imputed persons and some overseas visitors. Imputed persons are people known to exist but for whom no Census form was returned and so a statistical method was used to impute their demographic information. Overseas visitors are excluded from linking because they are not relevant in this context. These were people who indicated they usually lived in another country and expected to stay in Australia for less than a year.

Of the 20,928,304 Census records, 1,871,957 indicated they were born overseas with a year of arrival between 2000 and 2011. Of these, 2,787 were either born on Norfolk Island or inadequately described their country of birth, and 483,395 were born in New Zealand, of whom many would be non-visa settlers and thus not present on the SDB. A further 427,167 Census records who were either born overseas or did not provide a birthplace did not state their year of arrival.

4. THE LINKING PROCESS

This section provides an overview of the methodology employed to link datasets at the ABS, before proceeding to a detailed account of the method applied to create the SDB–Census linked file.

4.1 Linking methodology

The SDB and the 2011 Census datasets were linked using a probabilistic linking approach. The method aims to link records on two datasets (File A and File B), which belong to the same individual without a unique record identifier. Instead, records from the two files are linked using a number of variables common to both files. A key feature of this methodology is the ability to utilise a range of linking variables and record comparison methods to produce a single numeric measure of the likelihood two particular records belong to the same person. A record pair may be linked in spite of missing or disagreeing values on any given linking variable(s), providing there is sufficient agreement on other linking variables. The ABS uses a modified version of the open source data linking software Febrl (Christen, Churches and Hegland, 2004).

The probabilistic data linking process can be deconstructed into (i) standardisation, (ii) blocking, (iii) record pair comparison, and (iv) a decision model. These phases are outlined below.

4.1.1 Standardisation

Before records on the two datasets are compared, the contents of the two datasets need to be standardised to facilitate comparison. This includes a number of steps such as verification, recoding and reformatting fields, and parsing text fields. Additionally, some fields require substantial repair. For instance, a first name field may undergo a number of operations, such as the removal or recoding of non-alphabetic characters (hyphens and some blank spaces excepted), search and removal of common prefixes (Mr, Ms) and suffixes (Jr). Names may also undergo nickname standardisation or indexing to a name dictionary.

Some variables are coded differently at different points in time and concordances may be necessary to create variables which align on the two datasets. Variables may also be recoded or aggregated in order to obtain a more robust form of the variable. For example, field of qualification is coded at a fine level on Census data, but may not be reported consistently at this fine level. This means a new variable capturing job qualification at a broader level may be required for use in data linking.

This set of procedures is collectively termed *standardisation*. Standardisation takes place in conjunction with a broader evaluation of the dataset in which potential linking variables are identified.

4.1.2 Blocking

Once data files have been standardised, record pairs, consisting of one record from each file, can be compared to see whether they are likely to be a match, i.e. belong to the same person. However, if the files are even moderately large, comparing every record on File A with every record on File B is computationally infeasible. Blocking reduces the number of comparisons needed by only comparing record pairs where matches are likely to be found – namely, records which agree on a set of blocking variables. Blocking variables are selected based on their reliability and discriminatory power. Sex is only partially useful as it is typically well-reported, however it is minimally informative as it only divides datasets into two blocks, and is thus used in conjunction with other variables.

The process of blocking reduces the computational intensity of data linking. However, comparing only records that agree on a particular set of blocking variables means a record will not be compared with its match if it contains missing or poorly reported information on a blocking variable. To mitigate this, the linking process is repeated a number of times, using a range of different blocking strategies. For example, we may first block by Date of Birth and Sex, meaning we compare only records which agree on these variables. Records which we fail to link will proceed to the next blocking pass in which a different set of blocking variables are used. A linking strategy may consist of several *blocking passes* – that is, we may first block by Suburb and then subsequently block by Date of Birth. In this way, if a record has missing or inaccurate information or a legitimate change in value for a blocking variable the individual may still be linked in a later blocking pass using different blocking variables.

4.1.3 Record pair comparisons

Within a blocking pass, records on the two files which agree on the specified blocking variables are compared on a number of linking fields. Each linking field has associated *field weights* which are calculated prior to comparison. These reflect the amount of information that agreement, disagreement, or missing values, in a linking field provide about whether the records belong to the same individual). Field weights are based on two probabilities associated with each linking field: first, the probability that the field values agree on a record pair given that the two records belong to the same individual; and second, the probability that the field values agree on a record pair given the two records belong to different individuals. These are called m and u probabilities, or *match* and *unmatch* probabilities, respectively and are defined as:

$$m = P\left(\text{field agrees} \mid \text{records belong to the same individual}\right)$$
$$u = P\left(\text{field agrees} \mid \text{records belong to different individuals}\right)$$

Given that the *m* and *u* probabilities require knowledge of the true match status of record pairs, they cannot be known exactly, but rather must be estimated. The ABS uses a number of techniques to estimate m and u probabilities. For the series of 2011 linking projects, the ABS developed methodology to use the expectation maximisation (EM) algorithm to estimate these probabilities (see Samuels, 2011). In some instances the EM algorithm is deemed unsuitable, or fails to converge on an estimate, and in such cases *m* and *u* probabilities are based on those of similar linking projects.

As a new feature to the suite of 2011 Census Data Enhancement projects, *m* and *u* probabilities for missing data on a linking field were also calculated. These capture the probability that a linking field is missing on either or both datasets given the record pair belongs to the same individual (m_{missing}), or to two different individuals (u_{missing}). Note that *m* and *u* probabilities are calculated conditional on agreement on the specified blocking fields, as all records compared will agree on blocking variables.

The match (*m*) and unmatch (*u*) probabilities are then converted to agreement, disagreement and missing field weights. The formulae to convert *m* and *u* probabilities to field weights are a small extension of the Fellegi and Sunter (1969) linking methodology which now provide weights for missing data. They are as follows:

$$\text{Agreement weight } = \log_2\left(\frac{m}{u}\right), \tag{4.1}$$

$$\text{Missing weight } = \log_2\left(\frac{m_{\text{missing}}}{u_{\text{missing}}}\right), \tag{4.2}$$

and

$$\text{Disagreement weight } = \log_2\left(\frac{1-m-m_{\text{missing}}}{1-u-u_{\text{missing}}}\right). \tag{4.3}$$

Equations 4.1 to 4.3 give rise to a number of intuitive properties of the Fellegi–Sunter framework. First, in practice, agreement weights are always positive and disagreement weights are always negative. Second, the magnitude of the agreement weight is driven primarily by the likelihood of chance agreement. That is, a low probability of two random people agreeing on a field (e.g. Date of birth) will result in a large agreement weight applied when two records do agree. The magnitude of the disagreement weight is driven by the stability and reliability of a variable. That is, if a variable is well-reported and stable over time (e.g. Sex) then disagreement on the variable will yield a large negative weight. For each record pair comparison, the field weights from each linking field are summed to form an overall record pair comparison weight.

Before calculating m and u probabilities it is necessary for some variables to first define what constitutes agreement. Typical comparison functions include:

- Exact match (e.g. Sex). Agreement occurs only when the two field values are identical. This criterion is used for most linking fields.
- Approximate string comparison (e.g. Name). Two strings may be said to agree despite a certain proportion of missing, differing, or transposed characters, allowing for misspellings, transcriptions of poor handwriting, etc. String comparators (such as Jaro, Winkler comparator) can be used to ensure that both identical and similar string pairs are defined to agree.
- Numeric difference (e.g. Year of arrival). A pair may be defined to agree if their field values differ by an amount less than or equal to a specified maximum difference.

Alternatively, near or partial agreement may be factored into the linking process not in defining agreement but in converting m and u probabilities to weights. For example, in comparing the Year of arrival of migrants we may define agreement to be an exact match, and calculate m and u probabilities based on this definition, while additionally applying a partial agreement weight to arrivals which differ by one year.

Blocking fields, linking fields, agreement comparison functions, and m and u probabilities are passed into linking software. The records which agree on the blocking variables are compared on all linking fields.

4.1.4 Decision model

Finally, a decision model determines whether the record pair is linked, not linked or considered further as a possible link. The first phase of this process is automated, in which a record is assigned to its best possible pairing. This process is known as one-to-one assignment. Ideally, as is often true in practice, each record on File A has a single, obvious best pairing with a record on File B, which is its true match. Linking projects in the ABS have typically used an auction algorithm to optimally assign one record on File A to one record on File B. The auction algorithm maximises the sum of all the record pair comparison weights through alternative assignment choices. If a record A1 on File A links well to records B1 and B2 on File B, but record A2 links well to B2 only, the auction algorithm will assign A1 to B1 and A2 to B2, to maximise the overall comparison weights for all record pairs.

The second phase of the decision model stage takes the output of the one-to-one assignment and decides which pairs should be retained as links, and which should be rejected as non-links. This is done by defining cut-off weights against which record pair comparison weights are evaluated. The simplest decision rule uses a single cut-off such that all record pairs with a weight greater than or equal to the cut-off are assigned as links, and all those pairs with a weight less than the cut-off are assigned as non-links. A more sophisticated decision rule employs lower and upper cut-off weights. Record pairs with a weight above the upper cut-off are declared links while those with a weight below the lower cut-off are declared non-links. The record pairs with weights between the upper and lower cut-off weights are not automatically assigned a status, but designated for clerical review.

In clerical review, each record pair is manually inspected to resolve its match status. A clerical reviewer is often able to utilise information which cannot be captured in the automated comparison process, such as variations in names and common transcription errors (e.g. 1 and 7). Reviewed records are either accepted as links or rejected as non-links.

4.1.5 Multiple passes

As outlined in Section 4.1.2, the linking process is repeated a number of times, where each iteration uses a different set of blocking and linking variables. The use of a range of blocking strategies aims to ensure that records are compared with their true match at some point in the linking process. At the end of each blocking pass, the links are assessed, and this assessment can help shape the linking strategy in subsequent passes. For example, if the current set of linked records has not linked a particular demographic, a blocking pass can be constructed specifically to target this subpopulation.

When all blocking passes have been completed, the properties and quality of the linked dataset are assessed. Quality assessment includes estimation of the proportion of false and missed links, identification of over- and under-represented demographics, and an assessment of the behaviour of the linked dataset when used in analysis, such as regression.

4.2 Implementation in the Migrants Quality Study

The 2006 Migrants Quality Study (see Section 2) led to a number of recommendations for future SDB–Census linking. The ABS worked closely with DIAC to identify and resolve issues pertaining to the quality of the SDB data prior to the 2011 Census, and DIAC initiated strategies to address the identified issues prior to providing the final SDB file. This section provides an overview of the 2011 Migrants Quality Study. It outlines the Gold and Bronze Standard linking strategies, focussing on the measures adopted to resolve issues outlined in Section 2.

4.2.1 Candidate linking variables

Candidate linking variables (in addition to name and address) were identified using the following criteria. First, and most simply, the variable must appear in a similar form on the SDB and Census. Second, the variable ought to be applicable and non-missing for the population common to both datasets. Third, the variable ought to be well-reported and stable over time. Variables not meeting these criteria may still be used in probabilistic linking, but are less informative for identifying matches.

Table 4.1 shows missing rates for key candidate linking variables.

4.1 Missing data rates on candidate Bronze linking variables

	SDB missing		*Census missing*	
Variable	*Count*	*(%)*	*Count*	*(%)*
Age	1	0.00	109,332	0.52
Day of birth	1	0.00	2,008,275	9.60
Month of birth	1	0.00	2,009,294	9.60
Year of birth	1	0.00	98,834	0.47
Sex	32	0.00	291,627	1.39
Country of birth	2,725	0.21	626,072	2.99
Marital status*	422,745	32.14	4,179,771	19.97
Year of arrival**	65,535	4.98	15,874,099	74.85
Religion	977,424	74.32	1,434,517	6.85

* Includes “not applicable” – persons aged under 15
** Includes “not applicable” – persons born in Australia

4.2.2 Standardisation

The ABS has undertaken considerable work developing standardisation rules prior to linking, and much of this work was applied to the SDB–Census linking project. In particular, extensive work was done in standardising name and address fields. This section will briefly outline the key standardisation rules pertinent to the SDB–Census linking project.

Name

The first name field underwent substantial repair. In an iterative process, aliases in brackets were identified and removed, illegal characters removed, single character strings concatenated, common prefixes (titles) and suffixes removed, and both "full" first name (first name including middle names) and first name only fields were created. This constituted the repair process, and a similar set of rules were applied to family name. Standardisation took the repaired names as input and mapped them to a name dictionary.

It was noted that of the small percentage of records with missing first names on the SDB, many had multiple strings in the family name field. Investigations suggested that often the full name appeared in this field. Well-reported name fields in the SDB were used to create both first name and family name indexes, and these indexes were used to identify first names from the family name field.

Address

Address was parsed and the resulting address fields were used to repair and corroborate one another in an iterative process. For example, Suburb and Postcode were used to corroborate Street Name, and Street Name and Postcode used to corroborate Suburb. At the end of this process, addresses were geocoded.

Other variables

For the purpose of blocking and linking, some variables were collapsed into broader categories. Religion, originally coded on the Census to a four-digit level, was collapsed to a three-digit level, as the improved reliability outweighed the loss of discriminatory power from having fewer categories. Country of Birth was coded to both two- and four-digit levels, allowing the variable to be utilised in linking at both a both broad and fine level.

4.2.3 Blocking and linking strategy

The outcomes of the 2006 Migrants Quality Study, along with the preliminary analysis and standardisation of the files, led to the development of a new blocking and linking strategy in 2011. Table 4.2 outlines the Gold Standard blocking and linking strategy, and table 4.3 the Bronze Standard. Some linking strategy decisions were made in advance, while others were outcomes of analysis of the linking passes completed up to that point. This section describes the strategies, focussing in particular on the more substantial components introduced for the first time in this study.

4.2 Gold Standard blocking and linking strategy

Variable	*Tolerance*	*Pass 0*	*Pass 1*	*Pass 2*	*Pass 3*	*Pass 4*	*Pass 5*	*Pass 6*	*Pass 7*
Name information									
First name (standardised)	Winkler (0.8–0.85)	L							
First name (standardised)	Winkler (0.8–0.9)		L						
First name (standardised)	Winkler (0.9)			L					
First name (cleaned)	Winkler (0.84)						L		
First name (cleaned)	Winkler (0.85)				L	L		L	L
Last name	Winkler (0.8–0.85)	L							
Last name	Winkler (0.84)						L		
Last name	Winkler (0.85)				L	L		L	L
Last name	Winkler (0.9)			L					
First name initial (1st 2 chars of Fn)	Exact string			**B**					
Last name initial (1st char. of Sn)	Exact string						**B**		
Personal characteristics									
Date of Birth	Exact string	**B**							
Age	Exact numeric		L			L	**B**	**B**	L
Day of birth	Exact string		L	L	L	**B**	**B**	**B**	L
Month of birth	Exact string		L	L	L	**B**	**B**	**B**	L
Year of birth	Exact numeric			L	L				
Sex	Exact string		L	**B**	**B**	**B**	**B**	**B**	L
Country of birth (4 digit)	Exact string		L	L	L	L	L	L	L
Country of birth (2 digit)	Exact string	**B**							
Marital status	Exact string	L	L			L	L	L	L
Year of arrival	Exact string	L	L						L
Religion (3 digit)	Exact string	L							
Address information									
Street number	Exact string						L		
Street name	Exact string					L			
Street name	Winkler (0.9)			L					
Street name	Winkler (0.85)				L		L		
SA1	Exact string				**B**				
Postcode	Exact string					**B**			
Suburb	Winkler (0.85)	L							
Street name (Census 1 or 5 years ago)	Winkler (0.85)							L	
Street name (Census 1 or 5 years ago)	Winkler (0.9)								L
State (Census 1 or 5 years ago)	Exact string							L	L
Family indicator	Exact string		**B**						**B**

B : Blocking variable.
L : Linking variable.

A feature new to the 2011 Migrants Quality Study was the use of a family indicator as a blocking variable. Both the SDB and Census datasets contain a family indicator variable, and in order to create a concordance of these indicators, the Gold Standard commenced with an initial pass which aimed to link one member of every SDB family to its Census record. This initial linking run was termed *Pass 0* as the results did not contribute to the final linked dataset but instead were used to create an index of SDB and Census Family IDs. Pass 0 links contributed to this index provided they were of sufficient quality, and the index was used to merge the Census Family ID onto the SDB. This ID was used as a blocking variable in Pass 1, as it divided the files into many small blocks reducing comparison time. In Pass 7, very high-quality Family IDs were again used as a blocking variable, with the aim of linking children, who typically have fewer available linking variables.

The Gold Standard used a range of comparison methods. Names were compared using three different Winkler score thresholds, and additionally using an interpolation function in which names with a Winkler score between a specified upper and lower threshold were given a partial agreement weight. Further, linking passes used one of two first name alternatives, initially a name which had been standardised against a name index, and later a name which had undergone cleaning and repair but remained in an unstandardised form. This method was developed to help overcome errors in name data.

Despite the use of Medicare records to update SDB address prior to linking, clerical review of early linking passes identified a number of matches of moderate score which did not match on address. However, they did match on a Census previous address. Accordingly, Census previous address fields were developed for linking as follows. Previous address was set to Census address five years ago when present; if the field was missing, it was set to Census address one year ago, and if this too was missing, was set to current Census address. Pass 6 linked SDB address fields to Census previous address and in this way found SDB records whose address data had not been updated.

The Bronze Standard blocking and linking strategy adhered to the Gold Standard where possible, but with less available linking information was more straightforward. A family indicator variable was created in a manner similar to that described for the Gold Standard. However, the Bronze Family index was created not as a separate linking pass but as a by-product of Pass 1. Bronze Family ID was used as a blocking field in Pass 4, again with the aim of identifying children, who often have fewer well-reported applicable linking variables.

4.3 Bronze Standard blocking and linking strategy

Variable	*Tolerance*	*Pass 1*	*Pass 2*	*Pass 3*	*Pass 4*	*Pass 5*
Personal characteristics						
Date of Birth	Exact string		**B**	**B**		**B**
Age	Exact					
Day of birth	Exact string	L			L	
Month of birth	Exact string	L			L	
Year of birth	Exact	L			L	
Sex	Exact string	**B**	L	**B**	L	**B**
Country of birth (4 digit)	Exact string	L			L	L
Country of birth (2 digit)	Exact string		**B**			
Marital status	Exact string		L	L	L	L
Year of arrival	Exact string	L	L	L	L	
Year of arrival	Approx. numeric (±1)					L
Religion (3 digit)	Exact string		L		L	L
Address information						
Mesh Block	Exact	**B**		L		
SA1	Exact		**B**			L
Family indicator	Exact				**B**	

B : Blocking variable.
L : Linking variable.

Census address information permissible for use in the Bronze Standard is limited to Mesh Block, Statistical Area 1 (SA1), along with larger areas (SA2, SA3, SA4, State/Territory) which were not used in this study. These variables were not created for Census previous address (address one and five years ago), and there was accordingly no opportunity to conduct a Bronze equivalent of Gold Pass 6.

The omission of name and address variables means the Bronze Standard is accordingly more reliant on the remaining linking variables. There is little which can be done to overcome this, and the adopted blocking and linking strategy utilised the auxiliary variables Religion (an optional response variable on the Census) and Year of Arrival. Year of Arrival had been noted as problematic for onshore applicants, as the variable on the SDB refers to the year a migrant obtained their visa, whereas the Census collects the year a migrant first arrived in the country. To help mitigate this, both exact and approximate numeric comparators were employed for this variable in the Bronze Standards.

4.2.4 Cut-off weights and clerical review

As discussed in Section 4.1, record pairs are clerically reviewed at the end of each linking pass in order to ascertain the optimal cut-off weights to accept pairs as matches or reject pairs as non-matches. Across the seven passes of the Gold Standard, approximately 6,000 records were clerically reviewed in order to determine where to set upper and lower cut-off weights, and a further 10,000 falling between these cut-off weights were reviewed to determine their match status. The Bronze Standard used a single cut-off in each pass, which were set utilising knowledge from the Gold Standard exercise.

Three separate Bronze Standard linkages were conducted, each distinguished by the cut-off weight used; the three linked datasets being Bronze High, Bronze Medium, and Bronze Low. These variants were created to assess the impact that the level of the cut-off weight has on analysis of the linked data. Further, it is feasible that no single Bronze dataset should be optimal for all forms of analysis. Some applications may require a highly accurate linked dataset, with very few false links, whereas others will benefit from a dataset which is representative of the migrant population. In such cases, representative false links may be less damaging to the analytical goals than missing a difficult to link demographic.

Bronze High is comprised of the first two Bronze linking passes only. These passes accepted only high quality links; subsequently as true matches became more difficult to come by, batches of a lower quality were accepted. Bronze Medium was created using all five Bronze linking passes, and Bronze Low reviewed the five linking passes, in particular Pass 1 and Pass 2, in order to add batches of moderate quality which had previously been overlooked. The proportion of SDB records linked on both the Gold and Bronze Standards, along with the match-rate and link accuracy of all Bronze datasets, assessed against the Gold Standard, are discussed in Section 5.

5. EVALUATION OF THE LINKAGE

The linked datasets were evaluated on a number of measures. For this Quality Study, the authors considered the following:

- The number of links against the expected number of links;
- The properties of the Migrants Settlement Database (SDB) records that did not get linked to a Census record in the Gold Standard;
- The match-link rate and link accuracy of the different Bronze Standard linkages, using the Gold Standard as a benchmark;
- The under- or over-representation of sub-groups in the linked datasets compared with the Gold Standard and the SDB; and
- The effects of this under- or over-representation on some typical analyses and models fitted to linked data.

The Gold Standard is evaluated primarily using the first two criteria while the final three criteria evaluate Bronze Standards using the Gold Standard as a benchmark. Analysis also evaluated under- or over-representation of sub-groups on the Gold Standard against the SDB. This section provides a summary of the investigations performed.

5.1 Comparing expected number of links to actual number of links

Initially, it is important to consider how many SDB records we might reasonably expect to link to the Census. Persons on the SDB file might be missing from the Census file for several reasons, including:

- They were missed by the Census, thus contributing to Census undercount;
- They were temporarily out of the country on Census night;
- They emigrated from Australia before the Census; or
- They died since arriving in Australia.

As discussed in Section 3.1, persons on the SDB file who had departed Australia before Census night, either temporarily or permanently, were removed from the SDB prior to linking. Thus, we would expect the main causes of persons on the SDB file missing from the Census file to be that they were missed by the Census or had died since arriving in Australia.

The Census Post Enumeration Survey (PES) estimates that 507,573 persons born overseas who were in the country on Census night were not counted in the Census Country of Birth figures (ABS, 2012a). However, the majority of these persons provided a Census form and merely omitted a Country of Birth, and were therefore in scope to be linked. An estimated 121,089 of these 507,573 persons did not fill out a Census form. Of these, 26,412 were estimated to have arrived since the year 2000, and were therefore on the SDB but out of scope to be linked. Additionally, applying mortality rates by age group to the SDB suggests that 12,248 persons on the SDB died prior to the 2011 Census. Excluding these groups leaves 1,276,338 SDB records expected to be in scope on the 2011 Census. The analysis below is given in terms of the raw SDB records unless adjusted figures are specified.

Table 5.1 shows the number of records linked in the Gold and the three Bronze Standards. Bronze Standard files have fewer links, in accordance with the less informative set of variables available to link on. Lowering the cut-off weights yields more links, in accordance with the theory outlined in Section 4.

5.1 Number of SDB records available for linking and the numbers linked for Gold Standard and each level of Bronze Standard

	Number of records linked			
SDB records	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
1,315,048	1,103,389	671,962	905,814	1,003,532

Note: Bronze High, Medium, and Low refer to the three Bronze Standard linkages; High, Medium, and Low indicate the level of cut-off weights used for each linkage.

Table 5.2 gives Gold Standard links classified by blocking pass, along with the raw and adjusted percentage linked. The Gold Standard blocking strategy can be found in table 4.1, in the previous section. As expected, the majority of links are created in the initial passes, which identify all record pairs with strong agreeing information. The use of Census previous address in Pass 6 found a group of record pairs with out of date SDB address, which had not been identified in earlier passes.

5.2 Number of SDB records linked at each linking pass for the Gold Standard

	Number of input records	*Records linked*		*Adjusted records linked (%)*
		Count	*(%)*	
Pass 0	1,315,048			
Pass 1	1,315,048	772,620	58.8	60.5
Pass 2	542,428	205,305	15.6	16.1
Pass 3	337,123	77,591	5.9	6.1
Pass 4	259,532	8,005	0.6	0.6
Pass 5	251,527	6,597	0.5	0.5
Pass 6	244,930	25,119	1.9	2.0
Pass 7	219,811	8,152	0.6	0.6
All passes	1,315,048	1,103,389	83.9	86.4

5.2 Unlinked SDB records

The main reasons for not linking records on the SDB to the Census are:

- The corresponding Census record does not exist (as discussed in Section 5.1); and
- The quality of data on either the SDB record or the corresponding Census record is too poor to facilitate linking.

Data quality can be affected by respondents not completing key questions or making errors in the information provided. This issue affects both the SDB and Census datasets. Table 5.3 presents missing SDB variable frequencies for the 211,659 SDB records which were not linked in the Gold Standard.

5.3 Missing SDB fields for SDB records unlinked in the Gold Standard

Linking variable	*Number of unlinked records missing variable information*	*Percentage of 211,659 unlinked SDB records*	*Percentage of 1,315,048 SDB records*
First name	1,759	0.8	0.1
Last name	105	0.0	0.0
Sex	12	0.0	0.0
Age	1	0.0	0.0
Date of birth	1	0.0	0.0
Country of birth	616	0.3	0.0
Year of arrival	0	0.0	0.0
Marital status (age ≥ 18 years)	33,431	18.5 (a)	3.1 (b)
Religion	152,128	71.9	11.6
Mesh Block	28,525	13.5	2.2
Street number	22,092	10.4	1.7
Street name	21,488	10.2	1.6
Suburb	19,940	9.4	1.5
Postcode	18,610	8.8	1.4

(a) Based on 180,294 unlinked SDB records of persons aged 18 years and over.
(b) Based on 1,082,370 SDB records of persons aged 18 years and over.

Approximately 113,000 SDB records contained incomplete address information, and of these 60% were missing on all address fields. A further 21% provided a complete address which could not be coded to a Mesh Block. Relatively few SDB records were missing on only one address field. The pattern of concurrent missing address fields suggests that multiple missing address fields may have been a factor in failing to link approximately 20,000 of the unlinked SDB records.

Although out of date SDB address remains a potential problem in the SDB–Census linking project, the extensive effort to update SDB address data, along with the utilisation of Census previous address in the linking process appears to have mitigated the impact to a large extent.

Table 5.4 shows rates of multiple missing values in Bronze linking fields for unlinked SDB records in the Gold and Bronze Standards. The use of name and address saw the Gold Standard less reliant on other auxiliary variables, and proportionally more records with multiple missing variables were linked in the Gold Standard. However, missing rates on individual variables in the SDB were generally low, and multiple SDB missing values do not appear to be problematic for the Bronze Standard linked files. Missing Census fields on the migrant population may still pose a problem.

5.4 Multiple missing Bronze linking fields for unlinked SDB records in the Gold and Bronze Standards

	Number of missing fields					
Dataset	*0*	*1*	*2*	*3*	*4+*	*Total*
Gold Unlinked	43,160	124,332	43,888	275	4	211,659
	20.9%	58.7%	20.7%	0.1%	0.0%	100%
Bronze High Unlinked	131,538	366,147	144,410	985	6	643,086
	20.5%	56.9%	22.5%	0.1%	0.0%	100%
Bronze Medium Unlinked	75,225	239,177	92,521	709	4	407,636
	18.5%	58.6%	22.7%	0.2%	0.0%	100%
Bronze Low Unlinked	50,855	178,698	79,694	663	4	309,918
	16.4%	57.7%	25.7%	0.2%	0.0%	100%
SDB	243,580	741,553	328,375	1,534	6	1,315,048
	18.5%	56.4%	25.0%	0.1%	0.0%	100%

Contributing fields: Day and Month of Birth, Age, Sex, Country of Birth, Marital Status, and Religion

It was noted that disproportionately large groups of migrants from some countries report the same Date of Birth. These dates appear to be allocated administratively rather than actual birth dates. For example, 33.5% of migrants from Afghanistan reported January 1st as their Date of Birth while a further 8.0% reported December 31st. Provided that those people consistently use the same Date of Birth, this will not prevent them being linked. However, for these people, Date of Birth has substantially less distinguishing power.

Some SDB records remained unlinked because the quality of information on the equivalent Census record was insufficient to establish enough agreement to assign a link.

In summary, there were 211,659 unlinked SDB records in the Gold Standard linkage. It is estimated that 38,660 of these were not linked because the equivalent Census record did not exist and 20,000 were not linked because of missing address information. A potentially large number remained unlinked because of out of date address information on the SDB or because of poor data quality on either SDB or Census records.

The issues above, which are outlined in some detail for the Gold Standard also relate to the Bronze Standard linkages. Indeed, as Bronze Standard datasets do not utilise name, and use only Mesh Block as a geographic identifier, they are in one sense more reliant on quality address information, which is necessary to code Mesh Blocks.

5.3 Match-link rate and link accuracy

Matches are defined as record pairs comprising of two records which belong to the same person. The match-link rate is the proportion of true matches which are present in the linked dataset. Link accuracy is defined as the proportion of records in a linked dataset which are matches. False links are links on a dataset which are not matches, while missed links are matches which were not linked and therefore do not appear in the linked dataset. In evaluating a Bronze Standard linked dataset we make the assumption that the Gold Standard is a complete set of matches; that is, it has correctly linked every individual who has a record on both the SDB and Census. Although this assumption doesn't hold in practice, the Gold Standard does represent the optimal linking quality given the available linking information. Manual clerical review, conducted as part of the linking process, indicated that 98.7% of the Gold Standard links are correct. Match-link rate and link accuracy (Figure 5.5, and Bishop and Khoo, 2007) were calculated for each linkage level of the Bronze Standard by comparing them with the Gold Standard.

5.5 Method of calculating Match-link rate and Link accuracy

		Match status from Gold Standard		
		Matches	Non-matches	
Link status from Bronze Standard	Links	True links	False links	Total links
	Non-links	Missed links (Falsely non-linked)	True non-links	
		Total matches		

$$\text{Match-link rate} = \frac{\text{true links}}{\text{total matches}} = \frac{\text{Bronze links where Bronze equals Gold}}{\text{Total Gold links}} \,.$$

$$\text{Link accuracy} = \frac{\text{true links}}{\text{total links}} = \frac{\text{Bronze links where Bronze equals Gold}}{\text{Total Bronze links}} \,.$$

Table 5.6 compares the Bronze Standard datasets with the Gold Standard, showing the true and false links, match-link rate and link accuracy. As expected, lowering the cut-off allows more true links to be found, increasing the match-link rate, at the expense of introducing more false links into the dataset, decreasing link accuracy. The trade-off between link accuracy and match-link rate is considered in Section 5.4 and Section 5.5.

5.6 True links, false links, match-link rate and link accuracy for Bronze datasets

	True links	*False links*	*Match-link rate*	*Link accuracy*
Bronze high				
Pass 1	640,213	19,443	58.02	97.05
Pass 2	11,938	368	1.08	97.00
Pass 3	.	.	.	.
Pass 4	.	.	.	.
Pass 5	.	.	.	.
Total	652,151	19,811	59.10	97.05
Bronze medium				
Pass 1	640,213	19,443	58.02	97.05
Pass 2	11,938	368	1.08	97.00
Pass 3	151,202	29,763	13.70	83.55
Pass 4	26,976	4,575	2.44	85.50
Pass 5	15,107	6,229	1.37	70.81
Total	845,436	60,378	76.62	93.33
Bronze low				
Pass 1	645,477	24,916	58.50	96.28
Pass 2	13,792	895	1.12	93.91
Pass 3	165,539	45,571	15.00	78.41
Pass 4	26,976	4,575	2.44	85.50
Pass 5	36,297	39,494	3.29	47.89
Total	888,081	115,451	80.49	88.50
Total Gold links:	1,103,389			

5.4 Under- or over-representation of sub-groups

Various migrant characteristics from the SDB dataset were compared with the Gold and Bronze Standard linked files. In particular the relative frequencies of various subpopulations were analysed. The analysis indicated that no major subpopulations were missed in great numbers in the linking process. However, some groups appear to be more difficult to link than others, resulting in a small degree of under-representation on the linked files. It was found that rates of under-representation on the Gold Standard file were highest for migrants in the following groups:

- Migrants who have never been married;
- Migrants on family visas, in particular:
  - Family migrants aged 15 to 24 years; and
  - Family migrants born in North Africa and the Middle East, South-East Asia and North-East Asia;
- Migrants on humanitarian visas, in particular:
  - Humanitarian migrants aged 15 to 34; and
  - Humanitarian migrants born in North Africa and the Middle East, and to a lesser extent, Sub-Saharan Africa and North-East Asia;
- Migrants who reported their religion as Islam.

Further details of the under-representation of subpopulations is presented in tables 5.7(a), (b), (c), (d), (e), (f) and (g) below, which provide a comparison of the relative frequencies of migrant characteristics for the SDB dataset and the Gold Standard, as well as including information on the three Bronze Standards.

5.7(a) Relative frequencies (%) in each Visa category, for Gold and Bronze Standard linked data compared with the SDB

Visa category	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Skilled	57.1	58.2	57.7	56.8	57.1
Family	32.0	31.4	33.5	32.2	32.1
Humanitarian	10.6	10.0	8.6	10.8	10.5
Other	0.4	0.4	0.2	0.3	0.3

5.7(b) Relative frequencies (%) in each Marital status category, for persons aged 18 and over, for Gold and Bronze Standard linked data compared with the SDB

Marital status	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Never married	37.7	35.6	32.5	33.5	35.0
Widowed	1.3	1.3	1.2	1.3	1.2
Divorced	1.6	1.5	1.5	1.4	1.4
Separated	0.6	0.6	0.6	0.6	0.6
Married	58.8	61.0	64.2	63.2	61.8

5.7(c) Relative frequencies (%) by Religion, for Gold and Bronze Standard linked data compared with the SDB

Religion	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Buddhism	6.8	6.7	7.4	6.9	7.2
Christianity	51.7	52.3	52.0	51.6	50.2
Hinduism	5.6	6.0	6.3	6.3	6.4
Islam	18.4	17.7	16.9	18.2	18.2
Judaism	0.4	0.4	0.4	0.3	0.4
Other religions	2.8	3.0	3.0	3.2	3.1
No religion	14.3	14.1	14.1	13.5	14.7

5.7(d) Relative frequencies (%) of Family migrants in each Age group, for Gold and Bronze Standard linked data compared with the SDB

Age group	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
0–14 years	7.3	7.1	6.9	7.7	7.4
15–24 years	11.4	10.8	10.0	11.4	11.6
25–34 years	35.9	36.1	37.9	36.8	36.9
35–44 years	24.2	24.5	24.2	23.6	23.6
45–54 years	9.6	9.8	9.4	9.4	9.5
55–64 years	5.5	5.6	5.6	5.4	5.4
65+years	6.0	6.1	6.1	5.7	5.6

5.7(e) Relative frequencies (%) of Family migrants by Region of birth, for Gold and Bronze Standard linked data compared with the SDB

Region of birth	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Oceania	3.0	2.9	2.7	2.9	2.9
NW Europe	17.9	18.5	18.1	15.6	15.3
South & East Europe	4.4	4.5	4.5	4.6	4.5
North Africa & Mid East	7.9	7.5	7.4	7.9	7.8
SE Asia	23.1	22.7	23.1	24.0	24.0
NE Asia	18.3	17.9	17.6	18.1	18.7
South & Central Asia	12.6	13.1	14.2	14.2	14.2
Americas	7.5	7.6	7.5	7.6	7.4
Sub-Saharan Africa	5.3	5.3	5.1	5.3	5.2

5.7(f) Relative frequencies (%) of Humanitarian migrants in each Age group, for Gold and Bronze Standard linked data compared with the SDB

Age group	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
0–14 years	20.7	21.3	22.0	22.0	21.2
15–24 years	24.3	23.8	21.8	23.7	23.6
25–34 years	21.6	20.5	20.1	20.3	20.9
35–44 years	16.6	16.9	17.7	16.7	16.9
45–54 years	9.8	10.3	10.6	10.0	10.1
55–64 years	4.4	4.7	5.0	4.7	4.6
65+years	2.7	2.7	2.9	2.6	2.7

5.7(g) Relative frequencies (%) of Humanitarian migrants by Region of birth, for Gold and Bronze Standard linked data compared with the SDB

Region of birth	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Oceania	0.3	0.3	0.2	0.3	0.3
NW Europe	0.2	0.2	0.2	0.2	0.2
South & East Europe	3.0	3.2	2.7	3.0	2.9
North Africa & Mid East	44.3	43.7	42.5	43.2	43.4
SE Asia	10.6	11.3	12.9	12.2	11.7
NE Asia	2.1	1.9	2.0	1.9	2.1
South & Central Asia	18.9	19.4	20.2	19.6	20.0
Americas	0.2	0.2	0.2	0.2	0.2
Sub-Saharan Africa	20.1	19.9	19.1	19.5	19.2

On the Gold Standard dataset, family and humanitarian visa holders are under-represented, while skilled visa holders are slightly over-represented. In contrast, when comparing the three Bronze Standards to the Gold Standard, skilled visa holders are under-represented. Furthermore, on the Bronze Medium and Bronze Low datasets, family and humanitarian visa holders are slightly over-represented compared to the Gold Standard.

When considering visa class and marital status, Bronze Low resembles the Gold Standard more closely than the other levels. However, when considering other variables including age, region of birth and religion, the extent of under- or over-representation is lowest for Bronze Medium. This higher level of agreement between the Gold Standard and the Bronze Standards with lower cut-offs can probably be explained by the fact that records that are harder to link due to missing linking variables are more likely to be linked when the cut-offs are lowered.

Considering tables 5.7 (a), (b), (c), (d), (e), (f) and (g), the differences observed between the SDB and the Gold Standard are to some extent reflective of broader trends in statistical monitoring. Those who are young and unmarried tend to be more mobile, making them both more difficult to enumerate and more difficult to link.

Another factor that may have influenced the under-representation of some groups is the extent to which the different migrant groups have contact with administrative systems that are used to update address information on the SDB. Family migrants, who upon arrival already have the support of family members already in Australia, may seek less government support and have reduced interaction with administrative systems. Although humanitarian migrants are likely to require more government support, they are also more likely to have a lower English proficiency which may result in poor data quality on their Census records. In addition, some migrants, depending on their visa arrangements, may not be immediately eligible for government support upon their arrival in Australia. In contrast, skilled migrants are somewhat over-represented on the Gold Standard. The factors responsible for this are unclear, though it may be that this group as a whole provides higher quality Census data.

For the three Bronze Standard datasets, the false and missed links were also analysed and table 5.8 compares the numbers of these across the three datasets.

5.8 False links and missed links for Bronze datasets

	Bronze High	*Bronze Medium*	*Bronze Low*
False links			
Records not linked in Gold Standard	9,067	30,763	58,863
Records linked differently in Bronze Standard	10,744	29,615	56,588
Total	19,811	60,430	115,451
Missed links	440,494	227,523	157,957

The false links were further classified as either records that had not been linked on the Gold Standard or records that were linked to a different record on the Bronze Standard than the record they were linked to on the Gold Standard. For all three datasets, the false links are divided approximately evenly into these two groups.

Analysis of the false links showed that most SDB records that were falsely linked were still linked to a Census record with similar characteristics. For example, 92.5% of the false links for the Bronze Low Standard agree on Age, while 97.4% have an age difference of no greater than five years. 99.4% agree on Sex, and of those that do not, 74.3% agree on Date and Country of Birth. In contrast, a small proportion of the false links appear to be individuals with very different characteristics, with a common Day/Month of Birth or Country of Birth.

The number of missed links decreased significantly as the cut-offs were lowered, as would be expected. There was no clear indication why the missed links were not linked on the Bronze Standard. However, analysis of migrant characteristics in the missed links datasets did find that the relative frequencies of some subpopulations differed from the Gold Standard. For example, the links missed in the Bronze Medium and Bronze Low Standards had a higher proportion of skilled migrants and a smaller proportion of humanitarian and family migrants when compared to the Gold Standard. Furthermore, all of the missed links datasets contain a larger proportion of migrants who arrived between 2000 and 2004 and a correspondingly smaller proportion of migrants who arrived between 2005 and 2011, indicating that more recent migrants were easier to link in the Bronze Standard. This analysis aligns with previous research, suggesting that a lower cut-off, reducing the number of missed links at the expense of introducing some additional false links, yields a dataset which more closely aligns with the Gold Standard.

5.5 Some typical analyses and the impact on conclusions from using different linkage standards

The linked SDB–Census file has the potential to answer a range of research questions. This section compares the performance of the Bronze Standard datasets in an area of research interest – employment status for migrants with respect to arrival cohorts, visa categories and a range of pertinent Census predictor variables. The same analysis was conducted using the Gold and Bronze Standard linked datasets, to assess the effect of different linkage standards on the conclusions drawn from the analysis.

Preliminary analysis of the three Bronze Standards indicated that the Bronze Medium and Bronze Low Standards were most similar to Gold Standard. To further investigate the differences between the Gold and Bronze Standards, a model of employment status (Employed, Unemployed or Not in the Labour Force) was fitted to each dataset, using visa category from the SDB and several Census variables as explanatory variables. The explanatory variables included in the model were

- Age;
- Sex;
- Registered marital status;
- Proficiency in spoken English;
- Year of arrival;
- Highest educational attainment;
- Student status (studying or not studying);
- Unpaid child caring responsibilities;
- Place of usual residence (urban, regional or remote); and
- Visa category (skilled, family, humanitarian, other) (from SDB).

The same explanatory variables were used for each model to enable comparison of the model results. While there was some difference in the estimated coefficients, and significance of coefficients, between the four models, there is very little difference in the predicted probabilities of employment status for the examples given below.

Example person 1

Table 5.9(a) lists the predicted probabilities of employment status for 30 year old male, who arrived in 2000 on a Family visa, speaks English only, has a Bachelor degree and is not currently studying, is married but has no unpaid child caring responsibilities, and lives in an urban area.

5.9(a) Predicted probabilities of employment status, by each linkage standard, for example person 1 as described above

Linkage standard	*Pr(employed)*	*Pr(unemployed)*	*Pr(not in the labour force)*
Gold	0.9464	0.0186	0.0350
Bronze High	0.9511	0.0161	0.0328
Bronze Medium	0.9480	0.0173	0.0347
Bronze Low	0.9462	0.0183	0.0355

Example person 2

Table 5.9(b) lists the predicted probabilities of employment status for 18 year old female, who arrived in 2010 on a Humanitarian visa, does not speak English well, highest educational attainment is year 8 or below, but is currently studying, has never been married and has no unpaid child caring responsibilities, and lives in an urban area.

5.9(b) Predicted probabilities of employment status, by each linkage standard, for example person 2 as described above

Linkage standard	*Pr(employed)*	*Pr(unemployed)*	*Pr(not in the labour force)*
Gold	0.0410	0.0844	0.8746
Bronze High	0.0391	0.0833	0.8776
Bronze Medium	0.0387	0.0829	0.8784
Bronze Low	0.0388	0.0820	0.8791

A simple summary statistic that compares the model parameters estimated from the Bronze Standard with those estimated from the Gold Standard is the *deviance of coefficients* as defined by Chipperfield (2009). This can be interpreted as the average difference between a Bronze regression coefficient and its corresponding Gold coefficient, where the difference is measured in terms of the number of standard errors of the Gold coefficient.

The deviance statistic was 2.84 for Bronze High, 2.13 for Bronze Medium and 1.74 for Bronze Low. This indicates the coefficients for Bronze Low are closest to the corresponding coefficients calculated on the Gold Standard. However, for example person 2, the predicted probabilities for Bronze High are the closest to the predicted probabilities for the Gold Standard.

5.6 Applications

The purpose of this paper is to describe and evaluate the linking methodology and resulting linked datasets used to append key immigration variables to migrant Census records. However, it is worthwhile briefly emphasising the utility of the linked dataset in yielding a better understanding of migrant populations and outcomes.

As a simple example, we can now assess socio-economic characteristics of migrants by visa category. Table 5.10 presents Socio Economic Indexes for Areas deciles (SEIFA, derived from Census data) for migrants by broad visa category. SEIFA utilises Census employment, education, occupation, income, and housing variables to produce four area level scores for all regions in Australia. Table 5.10 gives the SEIFA Index of Advantage and Disadvantage. It shows migrants who arrive on Humanitarian visas typically settle in lower scoring SEIFA areas, whereas those who arrive on Skilled visas are settled in higher scoring SEIFA areas. Migrants who arrive on Family visas are evenly distributed across SEIFA deciles.

5.10 Relative frequencies (%) in each SEIFA decile by Visa category for the Bronze Low dataset

	Visa category			
SEIFA decile	*Skilled*	*Family*	*Humanitarian*	*Other*
1	4.61	11.86	31.61	12.18
2	6.48	11.06	20.60	12.61
3	7.65	9.79	13.38	10.92
4	8.66	9.88	10.31	11.15
5	9.70	9.67	7.28	9.23
6	10.50	9.58	5.75	9.76
7	11.35	9.44	4.23	9.63
8	12.54	9.54	3.36	8.67
9	13.72	9.86	2.10	8.21
10	14.80	9.32	1.39	7.64

Similar analysis could be conducted on finer visa categories, contrast recent and more established migrants, or focus on a particular component of SEIFA, such as the Index of Economic Resources. A full consideration of the potential of the linked dataset is beyond the scope of this paper, however the range of Census variables set to be made available for analysis will facilitate for researchers and policy makers an unprecedented understanding of migrant settlement challenges and outcomes.

5.7 Evaluation summary

In summary, 211,659, or 16% of the SDB file could not be linked to the Census file in the Gold Standard, when name, address, Mesh Block and other variables were used for linking. Of these records, 38,660 can be explained by known reasons that a Census record would not exist. A further 20,000 are probably due to missing and incomplete responses on the SDB, particularly address fields. The remaining 172,949 are most likely due to out of date address information on the SDB and poor quality responses for some Census or SDB records.

There are a number of factors which may contribute to poor quality data on the SDB and Census. For some migrants, English proficiency may be a barrier to providing complete and correct information. Errors may also arise in offshore application data when a proxy provides information on behalf of the applicant. Scanning errors are a known problem in Census data.

The under- or over-representation of subpopulations means care should be taken when interpreting analyses about those groups. It may be possible to overcome this problem by calibrating/weighting the linked data to known population counts for these subpopulations.

The 2006 Quality Study recommended that for future statistical research, the Bronze Low dataset should be used (Wright, Bishop and Ayre, 2009). Extensive analysis of other linked data found that missed links are more problematic than false links in drawing conclusions from linked data, and accepting some false links to reduce the number of missed links, yields a more comprehensive and representative linked dataset (Bishop, 2009; Chipperfield, 2009). The analysis presented here drew similar conclusions – a concession in link accuracy in order to increase match-link rate produces a better quality linked dataset – and the Bronze Low linked dataset is recommended for use in analysis.

ACKNOWLEDGEMENTS

This paper was prepared by the Analytical Services Branch (ASB), with assistance from the National Migrant Statistics Unit (NMSU). NMSU work program is jointly funded by the Australian Bureau of Statistics and the Department of Immigration and Citizenship (DIAC).

A large team from ASB and NMSU worked in close collaboration to produce the linked files. The linking project was given considerable support by IT staff at the ABS.

The authors are indebted to Guangyu Zhang, who made a number of important contributions to the linking methodology early in the project, vigorously seeking means to improve on the 2006 Quality Study.

The authors would like to thank Chris Heywood-Smith, Andrew Middleton, Jenny Dobak, Anne Weckert, Phillip Gould and Peter Rossiter for their valuable assistance and comments in the preparation of this paper. Chris Heywood-Smith in particular contributed extensively to producing an early draft paper structure and the work was invaluable.

REFERENCES

Australian Bureau of Statistics (2005) *Census Data Enhancement – Statement of Intention*, (last viewed on 9 August 2013) <http://www.abs.gov.au/websitedbs/D3110124.NSF/f5c7b8fb229cf017ca256973001fecec/5812a287d6a2e78fca2571ee001a7a49!OpenDocument >

—— (2006a) *Enhancing the Population Census: Developing a Longitudinal View*, Discussion Paper, cat. no. 2060.0, ABS, Canberra.

—— (2006b) *Information Paper: Census Data Enhancement Project: An Update*, cat. no. 2062.0, ABS, Canberra.

—— (2010a) *Information Paper: Census Data Enhancement Project: An Update, October 2010*, cat. no. 2062.0, ABS, Canberra.

—— (2010b) *Settlement Outcomes for Humanitarian Program Migrants – Experimental Estimates from the Migrants Statistical Study*, cat. no. 3416.0, ABS, Canberra.

—— (2010c) *Economic Outcomes of Skilled Program Migrants – Experimental Estimates from the Migrants Statistical Study*, cat. no. 3416.0, ABS, Canberra.

—— (2012) *Census of Population and Housing – Details of Undercount, 2011*, cat. no. 2940.0, ABS, Canberra.

—— (2013) *Migrants – Experimental Estimates from the Migrants Statistical Study*, cat. no. 3416.0, ABS, Canberra.

Department of Immigration and Citizenship (2009) *Longitudinal Survey of Immigrants to Australia*, DIAC, Canberra, (last viewed on 9 August 2013) < http://www.immi.gov.au/media/research/lsia/ >

Bishop, G. (2009) "Assessing the Likely Quality of the Statistical Longitudinal Census Dataset", *Methodology Research Papers*, cat. no. 1351.0.55.026, Australian Bureau of Statistics, Canberra.

Bishop, G. and Khoo, J. (2007) "Methodology of Evaluating the Quality of Probabilistic Linking", *Methodology Research Papers*, cat. no. 1351.0.55.018, Australian Bureau of Statistics, Canberra.

Chipperfield, J.O. (2009) "Generalised Linear Models with Probabilistically Linked Data", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.098, Australian Bureau of Statistics, Canberra.

Christen, P., Churches, T., and Hegland, M. (2004) "Febrl – A Parallel Open Source Data Linkage System", *Proceedings of the Eighth Pacific–Asia Conference, PAKDD 2004, Sydney, Australia,* pp. 638–647.

Fellegi, I.P. and Sunter, A.B. (1969) "A Theory for Record Linkage", *Journal of the American Statistical Association,* 64(328), pp. 1183–1210.

Samuels, C. (2011) "Using the EM Algorithm to Estimate the Parameters of the Fellegi–Sunter Model for Data Linking", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.120, Australian Bureau of Statistics, Canberra.

Solon, R. and Bishop, G. (2009) "A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset", *Methodology Research Papers*, cat. no. 1351.0.55.025, Australian Bureau of Statistics, Canberra.

Wright, J.; Bishop, G. and Ayre, T. (2009) "Assessing the Quality of Linking Migrant Settlement Records to Census Data", *Methodology Research Papers*, cat. no. 1351.0.55.027, Australian Bureau of Statistics, Canberra.

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** The ABS website is the best place for data from our publications and information about the ABS.

LIBRARY A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au

Produced by the Australian Bureau of Statistics