1351.0.55.051

Australian Bureau of Statistics

Research Paper

Feasibility Study of Linking Migrant Settlement Records to Personal Income Tax Data

www.abs.gov.au

New Issue

Research Paper

Feasibility Study of Linking Migrant Settlement Records to Personal Income Tax Data

Laura Walsh and Anne Weckert

Australian Bureau of Statistics

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 22 AUG 2014

ABS Catalogue no. 1351.0.55.051

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper. For further information, please contact Ms Lisa Conolly, Director, Regional and Migrants Statistics National Centre, on Adelaide (08) 8237 7402.

FEASIBILITY STUDY OF LINKING MIGRANT SETTLEMENT RECORDS TO PERSONAL INCOME TAX DATA

Laura Walsh and Anne Weckert
Australian Bureau of Statistics

EXECUTIVE SUMMARY

In 2013, the Australian Bureau of Statistics (ABS) conducted a study to examine and assess the feasibility of linking unit record data from the Australian Taxation Office (ATO) Personal Income Tax records with the Australian Government's Settlement Database (SDB). This study was known as the Migrant Personal Income Tax (PIT) Data Integration (DI) project.

The Migrant PIT DI project was conducted with the aim of assessing the viability and quality of linking the SDB with the ATO PIT data.

The primary benefits of the project are:

1. Increasing the potential for use of administrative data sources such as the SDB and PIT data by the ABS for the purpose of statistical output;
2. Creating an enriched dataset for statistical and research purposes;
3. Possible new statistics on recent migrants able to be produced at a relatively low cost and without additional burden to providers (to inform policy debate, decision making and evaluation); and
4. Advancing the capability of the ABS as an Integrating Authority through demonstration of the feasibility and statistical value of linking of administrative data.

The first phase of the Migrant PIT DI project is the Migrant PIT Linkage Feasibility Study. The study linked the SDB to the PIT records using variables such as name, date of birth and address. Figures from the linked file were compared with other data sources with comparable data items published by the ABS to assess the quality of the links and the linkage rate. Relevant legislation and guidelines, including the *Privacy Act 1988* and the *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes* were adhered to, protecting the privacy of individuals on both datasets.

Around 70% (68.8% in 2009/10 and 69.0% in 2010/11) of records on the SDB considered to be 'in-scope' for the feasibility study linked to a record on the PIT file. Of the unlinked SDB records, it is likely that a high proportion of them are simply non-tax lodgers, and therefore a match to a PIT record is not expected. If it were

possible to quantify this and exclude these records from the linkage process, hypothetically the linkage rate would have been much closer to 100%.

The project demonstrated that linking the SDB to PIT is feasible and can provide useful information on the economic contribution of individuals who have migrated to Australia. The feasibility study also demonstrated that data linkage using a limited range of variables is possible.

The next step for this project is to conduct a statistical study and, following consultation with relevant stakeholders, it is hoped that this second phase of the project can be conducted in 2014.

This paper provides background to the Migrant PIT Linkage Feasibility Study, a brief description of the linking strategy and process, a discussion of the quality of the linking of the SDB records to the PIT records, provides recommendations for improvements to the linking, and explores potential for an analysis dataset.

ABBREVIATIONS

ABN	Australian Business Number
ABR	Australian Business Register
ABS	Australian Bureau of Statistics
ACMID	Australian Census and Migrants Integrated Dataset
AMEP	Adult Migration Education Program
ASB	Analytical Services Branch
ATO	Australian Taxation Office
CDE	Census Data Enhancement
DI	Data Integration
DIAC	Department of Immigration and Citizenship
DIBP	Department of Immigration and Border Protection
DSS	Department of Social Services
ERP	Estimated Resident Population
ESTFN	Encrypted Scrambled Tax File Number
ETFN	Encrypted Tax File Number
ICSE	Integrated Client Services Environment
NMSU	National Migrants Statistics Unit
NTDBD	National Tax Data and Business Demography
PAYG	Pay As You Go
PID	Person level identifier
PIT	Personal Income Tax
PITMID	Personal Income Tax and Migrants Integrated Dataset
SDB	Settlement Database
SLK	Statistical Linkage Key
STFN	Scrambled Tax File Number
TRIPS	Travel and Immigration Processing System

ACKNOWLEDGEMENTS

This paper was prepared by the ABS' National Migrant Statistics Unit (NMSU), with assistance from the Analytical Services Branch. The NMSU work program is jointly funded by the Australian Bureau of Statistics, the Department of Immigration and Border Protection (DIBP) and the Department of Social Services (DSS).

The ABS acknowledges the assistance provided to the Migrant PIT Linkage Feasibility Study by the Department of Immigration and Border Protection, the Department of Social Services and the Australian Taxation Office.

The results of this study are based, in part, on tax data supplied by the ATO to the ABS under the *Taxation Administration Act 1953*, which requires that such data is only used for the purpose of administering the *Census and Statistics Act 1905*. Any discussion of data limitations or weaknesses is in the context of using the data for statistical integration purposes, and is not related to the ability of the data to support the ATO's core operational requirements.

The confidentiality of data is protected by legislation. The *Census and Statistics Act 1905* and the *Privacy Act 1988* require that all information collected by the ABS remain confidential. Both these Acts ensure that data submitted to, or collected by, the ABS are not provided to anyone where those data can be used to identify an individual. All ABS staff, including temporary employees, are legally bound never to release personal information to any individual or organisation outside the ABS. In addition, comprehensive security arrangements are implemented in ABS computer systems. These include use of regularly changed passwords, access control and audit trails.

The authors would like to acknowledge the extensive contribution of Caroline Deans for her work on the project. The authors would also like to acknowledge the advice and assistance with the linking methodology provided by the Analytical Services Branch, in particular Paul Campbell and Gokay Saher.

The authors would also like to acknowledge Brendan Kelly, Cassandra Elliot, Ben Mulder and the Outposted Officers to the ATO for their assistance in various stages of the project.

Finally, the authors acknowledge Andrew Middleton, Karen Connaughton, Phillip Gould, Jenny Dobak, Chris Heywood-Smith, Peter Rossiter and Ben Ashley for their valuable assistance and comments in the preparation of this paper.

CONTENTS

FEASIBILITY STUDY OF LINKING MIGRANT SETTLEMENT RECORDS TO PERSONAL INCOME TAX DATA

Laura Walsh and Anne Weckert
Australian Bureau of Statistics

ABSTRACT

In 2013, the Australian Bureau of Statistics was provided with access to the Australian Taxation Office's Personal Income Tax unit record data to assess the feasibility of linking records from the Australian Government's Settlement Database to the Personal Income Tax unit record data. The study concluded that linking was feasible, provided name and address information could be used as linking variables, and that the linked dataset could provide useful information that no other data source could provide. This paper provides background to the feasibility study, a brief description of the linking strategy and process, and an assessment of the quality of the linking. The potential benefits of future linking projects are also discussed.

1. INTRODUCTION

The Migrant Personal Income Tax (PIT) Data Integration (DI) project is aimed at integrating unit record data from the Australian Taxation Office (ATO) Personal Income Tax records with the Australian Government's Settlement Database (SDB) to create a new dataset for statistical and research purposes. The aim is to bring these datasets together using deterministic and probabilistic linking techniques.

The ABS has strong safeguards in place to protect identifiable information such as name and address, and these have been independently audited. These safeguards are backed by legislation (the *Census and Statistics Act 1905* and the *Privacy Act 1988*). Only those staff that have a need to view identifiable information as part of their duties have access to it, and only for a limited period of time. No information is released by the ABS in such a way that identifiable information compiled through linking can be associated with a specific person. This prohibition on release of identifiable information resulting from linking of datasets by the ABS is absolute – extending to all other parts of Government as well as the business and research communities.

The Migrant PIT Linkage Feasibility Study aims to append some key variables of interest from the SDB to the PIT data for the recent permanent migrant population. These variables include visa class (Skilled, Humanitarian and Family visas), application status (primary or secondary applicant), country of birth, year of arrival and whether the application was processed onshore or offshore. The resulting linked dataset has been referred to as the Personal Income Tax and Migrants Integrated Dataset (PITMID).

The Migrant PIT DI Project has three phases:

Phase 1: Feasibility

The Migrant PIT Linkage Feasibility Study aims to assess the quality of linking individual PIT records for the financial years 2009/10 and 2010/11 to the individual records of permanent migrants (who arrived on or after 1 January 2000 on the SDB) without use of a unique record identifier.

Phase 2: Dissemination

If Phase 1 suggests the linked dataset is of sufficient quality, aggregate data from the experimental 2009/10 PITMID and 2010/11 PITMID may be disseminated. Subsequently, it is anticipated that finer level data may be made available via customised consultancies or via a flexible output vehicle such as TableBuilder.

Phase 3: Production

If custodians, stakeholders and clients are satisfied with the quality, usefulness and protections provided to the data released during Phase 2, consideration will be given to undertaking the SDB/PIT integration and dissemination on an annual basis.

Some potential additional benefits of an annual series would include:

- Cohort analysis – Comparisons of economic outcomes of different cohorts who have arrived under different policy and economic conditions;
- Longitudinal analysis – The resulting longitudinally linked dataset could be used to assess a range of questions, notably the relationship between a migrant's visa class and their post-arrival social and economic pathways and outcomes in terms of income. The potential for longitudinal analysis could lead to improved research into, and identification of, the causal factors underlying migrant settlement outcomes.

Section 2 introduces the datasets to be linked, before outlining the linking methodology applied. It also presents an evaluation of the linked dataset and possible changes for future iterations.

Section 3 provides an overview of the quality and usefulness of the PIT data and discusses its potential to be used for other linkage projects.

Section 4 details some of the potential new statistics that could be produced from the new linked dataset as well as discussing their usefulness against some of the key data needs outlined by the stakeholders before the commencement of this project.

2. MIGRANT PIT LINKAGE FEASIBILITY STUDY

2.1 The datasets

This section provides an overview of the two sources being linked, namely the Settlement Database (SDB) and the 2009/10 and 2010/11 Personal Income Tax (PIT) files.

2.1.1 Settlement Database

The Settlement Database (SDB) is compiled by the Australian Government from various departmental systems and a number of external sources, including Medicare Australia (DIAC 2013). The Department of Social Services (DSS) has custodianship of the database.

The SDB is a consolidated database of people who have been granted a permanent or a provisional/temporary visa (DIAC 2013). The SDB generally excludes temporary visa holders. However, there are some records for people on provisional visas.

For Settlement visas that were granted onshore (i.e. in Australia), the Arrival Date refers to the latest date of arrival prior to the grant of that visa.

For Settlement visas that were granted offshore (i.e. outside of Australia), the Arrival Date refers to the first date of arrival after the grant of that visa.

Extract 1: SDB Client file

The SDB Client file extract used in this feasibility study covered the period 1 January 2000 to 6 March 2013 and contained the records of 1,998,473 persons who, during that period, were granted visas to live permanently in Australia.

The SDB extract contained both demographic information including name, date of birth, sex, country of birth, Australian citizenship date and information pertaining to their migration event including visa subclass, applicant status (primary/secondary) and location of application (onshore/offshore).

Extract 2: SDB Address history

A supplementary address history extract which contained 4,168,920 address records corresponding to the 1,998,473 persons on the SDB Client file (Extract 1) was also supplied. The records covered the period from 1 January 2000 to 6 March 2013. The address information on the SDB is updated monthly via an administrative process run from Medicare Australia. Medicare data is used to update address fields. Therefore, if an individual does not notify Medicare of a change of address, then the SDB record is not updated unless notification is received from the client directly or an update is received from the Adult Migration Education Program (AMEP). The number of addresses per person ranged from 1 to more than 20.

Extract 3: TRIPS Name history

A TRIPS Name History extract was used to evaluate and repair name fields. Names on the SDB extract are correct to the date of extraction. Names are updated from a variety of sources including Medicare, AMEP and manual updates. Individuals who change their name, e.g. women who change their family name when marrying or divorcing, present an issue for linking.

The TRIPS name history extract contained 407,666 records relating to the first names and surnames of 347,086 persons. The number of name records per person ranged from 1 to more than 10.

Extract 4: TRIPS Arrival and departure

TRIPS data also indicated whether, and if so when, migrants had last departed the country.

The extract used was originally provided for use in the creation of the *Australian Census and Migrants Integrated Dataset (ACMID), 2011*. The extract contained a person's last movement direction (arrival or departure) and last movement date information as at Census night (9 August 2011). Permission was granted from DIBP for this data to be utilised for the Migrant PIT Data Integration feasibility study.

The scope of this extract differs from the data for the Migrant PIT Data Integration feasibility study. The extract only contained records up to Census night (9 August 2011) rather than up to the date of the SDB extract for this project (6 March 2013).

Future iterations of this project would benefit from TRIPS arrival and departure extracts matching the SDB on reference period and scope. Similarly, the TRIPS extract could be expanded to include more arrival and departure information for an individual than just the last recorded movement. For more information on these issues, see Section 2.5.1.

2.1.2 Personal Income Tax records

The PIT is compiled by the ATO and consists of the following three datasets:

- Name and Address register – A dataset of name, address, sex and date of birth of persons who have submitted a tax return.
- Client data file – A dataset created by the individual tax return forms completed by individuals.
- Individual Pay As You Go (PAYG) file – A database created using the employer supplied tax return information for its employees.

Existing use of aggregate PIT data can be seen in the following publications:

- *Estimates of Personal Income for Small Areas* (ABS 2013d)
- *Wage and Salary Earner Statistics for Small Areas* (ABS 2013e)

The Personal Income Tax data used for the Migrant PIT Linkage Feasibility Study contained all persons who had a tax return processed for either the 2009/10 or the 2010/11 reference period at the time of extraction.

The PIT data is sourced from the Australian Taxation Office (ATO). The data is then supplied to the Australian Statistician under the *Taxation Administration Act 1953* for the purposes of administering the *Census and Statistics Act 1905*. The data has been collected in compliance with Australian taxation laws. The unit record data was provided to the ABS for a variety of statistical purposes and so was not tailored specifically to this project.

Data provided to the ABS by the ATO are from taxation returns processed up to 16 months after the end of the financial year (i.e. returns processed up to 31 October 2010 for the financial year ending 30 June 2009). This dataset can be requested annually by the ABS.

Due to the identifying nature of the data it contains, access to all ATO datasets is strictly regulated by the ATO. Both the ATO and the ABS handle personal information contained in the data in accordance with the Australian Privacy Principles contained in the *Privacy Act 1988*.

The unit record PIT dataset contains a range of key data items such as income and tax deductions. It also contains auxiliary socio-demographic data items such as age, sex and birth year. Information on the statistics contained in the dataset is generally available through the ATO website or via a combination of data dictionary and tax return form information.

According to taxation laws, individuals whose income is below a certain threshold are not required to submit tax returns. However, amendments to the taxation laws can significantly alter the information that is required to be reported in the personal income tax returns and statistics derived from the PIT dataset will be influenced by tax regulation changes. For the variables included in this analysis, there were no changes between the 2009/10 and 2010/11 financial years. The tax-free threshold in the 2009/10 and 2010/11 financial years was $6,000.

Extract 5: Name and Address register

The ATO Name and Address register extracts contained demographic information of persons who have submitted a tax return. The number of records received for each financial year is as follows:

- 2009/10 extract contained 12,432,776 person records, and
- 2010/11 extract contained 12,725,423 person records.

The two files were merged using a common unique identifier, the Encrypted Scrambled Tax File Number (ESTFN). The merged file contained 13,630,483 person records.

- 11,527,716 person records were present on both files,
- 905,060 were only present on the 2009/10 file, and
- 1,197,707 were only present on the 2010/11 file.

Extract 6: Client data

The ATO Client data file contained a range of variables. These included:

- Wage and salary income;
- Own unincorporated business income (both primary and non-primary production);
- Investment income (including rental income);
- Superannuation and annuity income;
- Government pensions and allowances (including tax-free payments);
- Income generated from interest and dividends; and
- Foreign sources of income.

A typical person does not need to fill out every data item in a tax return. As a result, most records have the majority of all possible fields blank.

Extract 7: Individual Pay As You Go (PAYG)

The PAYG extracts contained wage and salary information from employers submitted to the ATO that corresponded with the person record on the client register extract.

Three analysis variables are available on the PAYG file. These are:

- Scrambled Tax File Number (STFN);
- Australian Business Number (ABN); and
- Gross salary payment.

The Australian Business Register (ABR) stores details about businesses by ABN. This information could be used to identify the industry activity of the businesses.

PIT data record counts

After the linking process was completed, Client data and PAYG file information was provided only for those records that linked to an SDB record for each reference period.

The PAYG file contained duplicate records where an individual worked in more than one job. The number of records for each individual on the file ranged from 1 to greater than 20 on both files.

2.1 Record counts for SDB integrated Client data file and PAYG file

	2009/10	*2010/11*
ATO PAYG records	1,032,913	1,158,611
ATO Client data records	812,482	888,542

There were a few records that did not appear in both datasets. The total number of records that were present in either the ATO Client data file or PAYG file or both was 812,506 in the 2009/10 file and 889,249 in the 2010/11 file.

See Section 4 for potential new statistics using these records.

2.2 The linking process

This section provides an overview of the methodology employed to link the Settlement Database (SDB) and the Personal Income Tax (PIT) datasets to create the Personal Income Tax and Migrants Integrated Dataset (PITMID) for the 2013 Migrant PIT Linkage Feasibility Study.

The linking process 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.

2.2.1 Deterministic vs probabilistic linking

During the planning stages of the project it was expected that a Probabilistic linking[1] methodology would be used to integrate the SDB and PIT datasets.

A Probabilistic linkage methodology had previously been used by the ABS to create the *Australian Census and Migrants Integrated Dataset (ACMID), 2011*. For the ACMID, 2011 Linking project, the SDB records were predominantly a subset of the

1 Probabilistic linking links two variables together using link weights to calculate the probability that two given records refer to the same entity.

Census of Population and Housing records. Consequently, in that linkage, most out of scope records were able to be removed prior to linking. Further details of that integration project can be found in Richter *et al.* (2013).

This meant that, in theory, after the removal of the out of scope records, 100% of the remaining SDB file should be present in the Census file. However, this is not the case with the 2013 Migrant PIT Linkage Feasibility Study as there is no way of removing those records on the SDB without a PIT record prior to linking.

After analysis of the SDB and PIT datasets, it was decided that the project would use a Deterministic approach.[2] Reasons for this included:

- Available variables were high quality, well-reported and stable and could accurately distinguish between people the majority of the time. Further, there was a high rate of completeness in the datasets for linking variables (see Appendix B).
- There was a tight timeframe for conducting the feasibility study. Deterministic linking is quicker to run and it does not require the calculation of *m* and *u* probabilities.[3]
- As neither population is a subset of the other, probabilistic linking could result in finding the closest match where, in fact, there may not be a true match between the two files, resulting in false links. False links are of particular concern in this project due to their enduring nature if a concordance file is used to create a longitudinal dataset (see figure 2.3 in Section 2.2.7).

In theory, Deterministic linking should produce the same result as Probabilistic linking, if stringent criteria for assigning links which agree on every linking variable are used.

Experience from the ACMID, 2011 Linking Project suggested that a large proportion of the SDB records that have a true match on the PIT files, would exactly match on all of the linking variables. This assumption, combined with the concerns discussed above, resulted in the decision to first employ a deterministic linking methodology followed by probabilistic linking passes. The criteria for variable agreement were gradually relaxed in a targeted manner. The high quality links created using the deterministic method and the first probabilistic run are referred to as "high" links, whilst links assigned through the final probabilistic run, which relaxed the linking criteria, are referred to as "low" links. See Section 2.3.1 for more information on "high" links and "low" links.

2 Deterministic linking links two records together if all linking fields are identical and only one such record meets this criterion.

3 An *m* probability is the probability that two fields agree if they belong to the same record. A *u* probability is the probability that two fields agree if they belong to different records.

This project was unprecedented in the ABS for the small overlap between the SDB and PIT datasets. Requiring a high level of evidence that two records belonged to the same person was essential to ensure high quality links.

2.2.2 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. 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. 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.

This feasibility study used the same SDB file as that utilised for the ACMID, 2011 Linking Project. For that project, routine work was done to standardise name and address fields on the SDB file. The standardisation processes applied to the Name and Address variables on the SDB and the PIT are outlined in more detail in Richter *et al.* (2013).

2.2.3 Deterministic linking

Deterministic linking compares two records on a set of variables. If all variables agree, they are considered a link. Multiple iterations, or passes, can be undertaken using different sets of variables on which to link.

Typically, deterministic linkage begins by using very stringent matching rules (where the records pairs need to agree exactly on as many linking fields as possible). As some matches will not agree on all linking variables, subsequent deterministic passes typically relax linking conditions by removing a variable from the set of comparison variables. However, in order to establish a unique agreement, all sets of variables used in deterministic passes must be strongly identifying. The strength of deterministic linking is that it can quickly locate the high quality matches in a dataset.

The deterministic linkage for this study utilised 20 variables from the SDB dataset and 15 variables from the PIT dataset. Some variants of the base variables were produced through the cleaning and standardisation process and were also used in the linking process. A total of 22 passes were undertaken to match record pairs across the two datasets using deterministic linking. Higher quality links were collected and removed from the record pool of potential links at the end of each successive pass.

Eight of the deterministic passes utilised the Jaro-Winkler distance[4] (Jaro, 1989) and the Levenshtein distance[5] (Levenshtein, 1966) string comparators to relax the definition of agreement for name fields, allowing small differences, which typically occur as a result of transcription or scanning errors, to be ignored.

The SDB–PIT linking process then progressed from a deterministic to a probabilistic linking approach which is outlined in the next section.

2.2.4 Probabilistic linking

A key feature of Probabilistic linking 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.

Two passes were run to match record pairs across the two datasets using Probabilistic linking.

The ABS Probabilistic linking process is outlined in more detail in Richter *et al.* (2013).

2.2.5 Candidate linking variables

The first step in the linking process was to identify candidate linking variables. The following criteria were used:

1. The variables must be comparable on both the SDB and PIT file.
2. The variable ought to be applicable and non-missing for the population common to both datasets.
3. The variable ought to be well-reported and stable over time.

Variables not meeting the second and third criteria could still be used in probabilistic linking, but were less informative for identifying matches.

Two standards for linking are mentioned in this paper: Gold standard and Bronze standard. Where name and address details are used as linking variables, the linkage is considered Gold standard. Where name and address are not utilised for linking, the linkage is referred to as Bronze standard.

4 The Jaro-Winkler distance is a measure of the similarity of two strings. It uses the number of matching characters and the number of transpositions to create a matching "score". The score is normalised such that 0 equates to no similarity and 1 is an exact match.

5 The Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. The distance ranges from 0 for an exact match and at most is the length of the longer string.

Due to the high quality name and address information available on both files and the limited number of other common variables, this feasibility study used the Gold standard. The variables used in the linking process are listed in table 2.2 below.

2.2 Variables used in linking

Variable type	*SDB*	*PIT*
Name information	Given names	First given name
		Other given names
	Surname	Surname or family name
Age-related Information	Date of birth	Date of birth
Personal characteristics	Sex	Sex
Address information	Current address line 1	Address line 1
	Current address line 2	Address line 2
	Current address line 3 (Suburb)	Address line 3 (Suburb)
	Current state	State
	Current postcode	Postcode
	–	Country
	Previous address line 1	–
	Previous address line 2	–
	Previous address line 3 (Suburb)	–
	Previous state	–
	Previous postcode	–
	Forward address line 1	–
	Forward address line 2	–
	Forward address line 3 (Suburb)	–
	Forward state	–
	Forward postcode	–
	All Street names from address history	
Spouse information	–	Spouse given name
	–	Spouse surname or family name
	–	Spouse date of birth
	–	Spouse sex

An analysis of the candidate linking variables indicated that there was a very low rate of missing data for most of the variables and there were no significant missing variables. For more details on the missing rates of the candidate linking variables, see Appendix B.

2.2.6 Blocking and linking strategy

This section describes the blocking and linking strategies employed, focussing on features not used in the previous linking project involving the SDB and the 2011 Census of Population and Housing and thus being introduced for the first time in this study. The linking process is repeated a number of times, where each iteration uses a different set of blocking and linking variables. At the end of each linkage 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 on a particular demographic, a linkage pass can be constructed specifically to target this subpopulation. Appendix A outlines the detailed blocking and linking strategy for each of the 24 passes.

Address information

A complete address history was provided for all records on the SDB file. The SDB address file contained 4,168,920 records that corresponded to the original 1,998,473 person records. Each person record had at least one address record, with another record completed for each address change. The highest number of addresses per person was over 20.

Where address information has been used for a pass, 3 passes were undertaken. Using the date the address became effective, the recorded address at the time of the PIT reference period was recorded as the current address. The first compared the PIT address variables to the SDB "current" address variables, the second compared the PIT address variables to the SDB "previous" address variables (where available), and the third compared the PIT address variables to the SDB "forwarding" address variables (where available).

Date of birth

Three birth dates appeared in the SDB with greater frequency than any other. These dates are recorded by a person who only provides their year of birth, i.e. the day and month are unknown. These dates are referred to as "administrative dates" for this linking project. These dates are:

- January 1st
- July 1st
- December 31st

For passes 14 and 18, these dates of birth were excluded.

Levenshtein distance

In passes 5 and 6, the Levenshtein distance was used on the middle name to allow a tolerance for error.

If both records had a middle name, the names were defined to agree if they had a Levenshtein distance less than or equal to two. As middle name is sometimes omitted on a form, a missing middle name was treated as agreeing with any middle name.

Spouse linking

Spouse information from the PIT file was used to further identify matches between the two files. A couple can be identified on the PIT file using the available spouse information. If the same two people could be identified on the SDB file, the address information for the two SDB records was analysed to determine whether they matched at some point in their address history. This was then taken as evidence of a match on the PIT file.

Probabilistic Passes

Two probabilistic passes were run. Pass 23 used a high acceptance threshold and pass 24 used a lower threshold. Approximately 6,000 records were clerically reviewed in order to determine where to set upper and lower cut-off weights.

In both probabilistic passes, records were reviewed to determine their match status based on their linking weight. They were subsequently given "high" link or "low" link status.

When all linkage passes had been completed, the properties and quality of the linked dataset was assessed. Quality assessment included estimation of the match rate, identification of over- and under-represented demographics and an assessment of possible improvements to the linking process. This process is discussed in Sections 2.3, 2.4 and 2.5.

2.2.7 Future linkage

In this feasibility study, the linkage process was run on the full SDB file. However if the study is repeated annually with a new wave of SDB and PIT records each year, the time and resources required to conduct the linkage has the potential to increase each year as well.

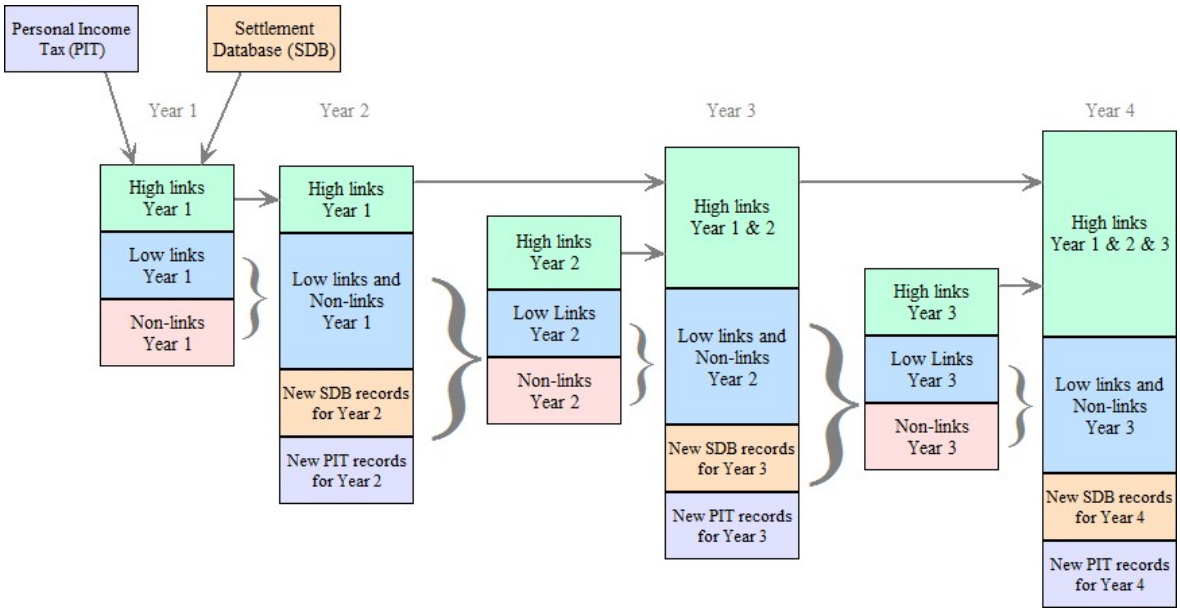
The linking methodology required to create an annual series could be modified to be somewhat simpler than that used to create the initial linked dataset. Records linked with a high degree of certainty can be used to create a concordance file of Migrant IDs and Encrypted Scrambled Tax File Numbers (ESTFN). Future waves of PIT data can initially be merged to the SDB using ESTFN and this concordance. The linking

process as applied in this feasibility study would then need to be applied only to records which had not previously linked with a high degree of certainty plus any new records that had been added to either file. The concordance file would be updated with new Migrant ID/ESTFN concordances each year as new links are discovered. This removes the potential problem of the dataset becoming cumulatively larger with each successive year.

This approach creates the opportunity to develop a longitudinal series as the SDB information for records linked with a high degree of certainty can be enhanced in each successive year with PIT data from the same linked SDB–PIT record using the concordance of Migrant IDs and ESTFNs.

Figure 2.3 illustrates this conceptual model for building up the selection of high quality links from year to year.

2.3 Conceptual model for building a longitudinal Migrant PIT integrated dataset



In addition, if a "low" link record links to the same PIT record with a "low" link the following year, then this record could be considered a high link from this point on and could be excluded from any future linking processes.

2.3 Evaluation of the linkage

The linked datasets were evaluated on a number of measures. For this feasibility study, the following were considered:

- The actual number of links against the number of links which might reasonably be expected;
- The properties of the Settlement Database (SDB) records that did not get linked to a PIT record;
- The match rate and link accuracy of the linkages, using the following datasets to calculate the expected number of matches:
    - *Australian Demographic Statistics* (ABS, 2010);
    - *Australian Census of Population and Housing, 2011* (ABS, 2012);
    - *Australian Census and Migrants Integrated Dataset, 2011* (ABS, 2014a).
- The under- or over-representation of sub-groups in the linked datasets; and
- Considerations for future iterations.

The linked dataset has been referred to as the Personal Income Tax and Migrants Integrated Dataset (PITMID).

2.3.1 Pass results

Of the 1,773,333 SDB records that went into the linking process, 951,234 (or 54%) were assigned a matching PIT record. Of these records, 786,977 (83% of the links) were assigned in the deterministic passes and the first probabilistic pass 23 (referred to in this paper as "high links"). A further 164,257 (17% of the links) were assigned in the second probabilistic pass 24 (referred to as "low links").

2.4 Number of linked SDB records by link method

	No. of links	*% of total links*
Deterministic	781,792	82.0
Probabilistic	169,442	18.0
Total	951,234	100.0

2.5 Number of linked SDB records by link type

	No. of records	*% of records*
High links	786,977	44.4
Low links	164,257	9.3
All links	951,234	53.6
No link	822,099	46.6
Total	1,773,333	100.0

2.3.2 Comparing expected number of links to actual number of links

The main reasons for not linking an SDB record to a PIT record are:

- There is no corresponding PIT record; or
- The data on either the SDB record or the corresponding PIT record is of insufficient quality to link.

Analysis of the SDB and the PIT datasets prior to linking indicated that they were both of a sufficiently high quality to link if a corresponding PIT record existed. However, there are a number of valid reasons why an SDB record would not have a corresponding PIT record. These include but are not restricted to a person:

- not having submitted a tax return by the time of extraction of information;
- not being in paid employment or earning an income in the reference period;
- living in Australia on a visa stream that places restrictions on employment;
- not being in Australia during the reference period;
- having an income that is within the tax free threshold; and
- not having any tax withheld.

As such, it was not expected that the linkage rate for this feasibility study would reach 100 per cent for the entire population on the SDB.

Therefore, in this feasibility phase, the 'in-scope' population was restricted by age, visa stream, date of permanent arrival in Australia and date of permanent departure.[6] The reasons for limiting the records by these variables are outlined below. The SDB records considered to be 'in-scope' were records that met all four of the following characteristics:

- Aged 15 to 64 years at the beginning of the reference period;
- Holding a Skilled, Family or Humanitarian visa;
- Permanent residence start date was before the start of the reference period; and
- Last movement was not a permanent departure prior to the set reference period.

In order to estimate the success rate of the linkage, it was considered desirable to narrow the analysis to a subset of people who are likely to have a corresponding PIT record, i.e. the in-scope population.

6 Additional data could be requested in future iterations of this project to account for permanent departures from Australia prior to and during the reference period. At the very minimum, by calculating the last date of departure at the end of the reference period, persons for whom that date was prior to the beginning of the reference period could be excluded. However, this does not mean that they did not earn a taxable Australian income while outside Australia during the reference period.

Although the scope limitations made the evaluation significantly more meaningful, it should be noted that removing these records did not definitively remove all of those out-of-scope people. A possibly large but unknown number of people remain who were not covered by these rules (and therefore were not deemed to be out-of-scope and removed). For this reason, if different rules were applied to determine the in-scope population even more accurately, then even higher linkage rates would be obtained than those presented in this feasibility study.

Age

Age was restricted to people aged between 15 and 64 years at the beginning of the reference period for in-scope records. This is generally considered to be the working age population according to the ABS *Labour Force Survey (LFS)* (ABS, 2014b). However, people outside of this age bracket do earn a taxable income and were linked to a PIT record but they were deemed to be out of scope for the purposes of this feasibility study (see Appendix C).

Visa stream

Several types of visas exist that have conditions that include restrictions on employment. These visas are mainly provisional visas. For this reason, the analysis was restricted to Permanent visa classes from the following visa streams:

- Skilled;
- Family; or
- Humanitarian.

While the main purpose of the SDB is to store information on permanent migrants, there are some records for people on provisional visas (DIAC, 2013). Records with these types of visas were excluded for the purpose of the study (see Appendix C).

Permanent residence start date

Migrants who have not been permanent residents of Australia for the entire reference period may have less of an opportunity to earn a taxable income in the reference period. This could be due to either not having arrived in Australia, or be living in Australia on a visa that places restrictions on employment.

For example, migrants who are not earning a taxable income for the whole reference period could distort results by registering a lower income than if they had been in-scope for the whole reference period. The limited time has the potential to place their income within the tax-free threshold and they may not have submitted a tax return.

For this reason, records were considered to be in-scope if their permanent residence start date was before the start of the reference period (e.g. 1 July 2009 for the reference period 2009/10). See Appendix C for the number of in-scope records for permanent residence start date.

This analysis uses the same definition of start date for permanent residence as the DIBP (DIAC 2013). DIBP calculates the start date of permanent residence differently for onshore and offshore applicants. For an onshore applicant, the start date for permanent residence is the date that the permanent visa was issued. It is worth noting that this means that a permanent onshore applicant may have been paying tax on a temporary visa prior to gaining permanent residence. For an offshore applicant it is the date of their first arrival in Australia since obtaining the permanent visa. This is illustrated in table 2.6.

2.6 Calculation of permanent residence start date

Name	*Date of first arrival*	*Location*	*Date of last entry*	*Visa issue date*	*Permanent residence start date*	*2009/10 reference period*	*2010/11 reference period*
Jane Citizen	01/06/2009	Offshore	NA	15/05/2009	01/06/2009	In-scope	In-scope
John Citizen	01/08/2009	Onshore	01/05/2010	01/06/2010	01/06/2010		In-scope

There were also over 500,000 people who were out-of-scope for permanent residence start date but in-scope for Age, Visa stream and departure information. Relaxing this condition to include these people may have an effect on the linkage rates and the distribution of income. Further analysis into the relaxing of permanent residence date could be considered in the future to find out whether this would have any significant impact.

In any future iterations of this project, an improvement in the linkage could be achieved by including data to account for permanent departures from Australia both 'prior to' and 'during' the reference period.

Departure date

People who departed Australia permanently prior to the reference period may have less of an opportunity to earn a taxable income during the same reference period. For this reason, in-scope records were limited to those people whose last movement was not a permanent departure prior to the reference period (see Appendix C). However, some people who departed Australia prior to the reference period had earned a taxable income and were linked to a PIT record. The economic impact of these offshore migrants does warrant further analysis.

In-scope SDB records

After all the scope restrictions were implemented, the remaining in-scope dataset had the following characteristics:

2.7 In-scope SDB records

	2009/10	*2010/11*
Number of in-scope SDB records	891,290	1,008,399
Number of in-scope SDB records that linked to a PIT record	613,584	695,480
Proportion of in-scope SDB records that linked to a PIT record	68.8%	69.0%

See Appendix C for a full discussion on how the final number of in-scope records was determined.

2.3.3 Characteristics of linked and unlinked SDB records

This section examines how the characteristics of linked and unlinked in-scope records differ according to variables of interest. These variables are analysed in the context of whether a person is a migrant from a main English-speaking country, their visa stream and their year of arrival.

Country of birth

Main English-speaking countries are the countries from which Australia receives migrants who are most likely to speak English. For the purposes of this project, main English-speaking countries have been defined as:

- The United Kingdom
- Ireland
- New Zealand
- Canada
- South Africa
- The United States of America

Records of persons from a main English-speaking country were more likely to link to a PIT record than records of persons from other countries.

For the 2009/10 reference period, 80% of SDB records of persons who indicated that they were born in a main English-speaking country were linked to a PIT record. This compared with 66% of records of persons from other countries who linked to a PIT record (table 2.8).

2.8 In-scope SDB records by Country of birth for 2009/10 reference period

Country of birth	Linked (No.)	Unlinked (No.)	Total (No.)	Linked (%)
Main English-speaking country	166,672	42,150	208,822	79.8
Other country	446,912	235,556	682,468	65.5
Total	613,584	277,706	891,290	68.8

For the 2010/11 reference period, 81% of SDB records of persons who indicated that they were born in a main English-speaking country were linked to a PIT record. This compared with 66% of records of persons from other countries who linked to a PIT record (table 2.9).

2.9 In-scope SDB records by Country of birth for 2010/11 reference period

Country of birth	Linked (No.)	Unlinked (No.)	Total (No.)	Linked (%)
Main English-speaking country	188,292	45,733	234,025	80.5
Other country	507,188	267,186	774,374	65.5
Total	695,480	312,919	1,008,399	69.0

Visa stream

When comparing main visa stream, people on a Skilled visa were most likely to link to a PIT record. For the 2009/10 reference period, 75% of person records on a Skilled visa linked, compared with 63% of people on a Family visa and 50% of people on a Humanitarian visa (table 2.10).

2.10 In-scope SDB records by Visa status for 2009/10 reference period

Visa stream	Linked (No.)	Unlinked (No.)	Total (No.)	Linked (%)
Skilled	382,873	125,332	508,205	75.3
Family	189,311	110,528	299,839	63.1
Humanitarian	41,400	41,846	83,246	49.7
Total	613,584	277,706	891,290	68.8

For the 2010/11 reference period, 76% of person records on a Skilled visa linked, compared with 63% of people on a Family visa and 48% of people on a Humanitarian visa (table 2.11).

2.11 In-scope SDB records by Visa status for 2010/11 reference period

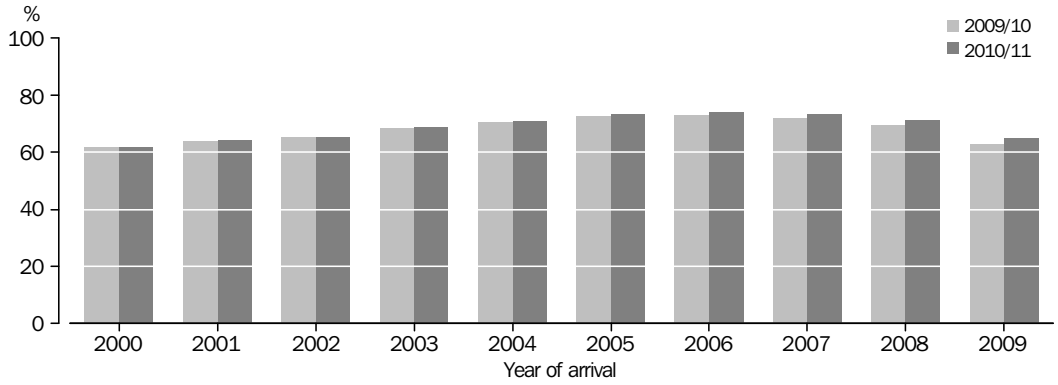
Visa stream	Linked (No.)	Unlinked (No.)	Total (No.)	Linked (%)
Skilled	436,277	137,321	573,598	76.1
Family	213,240	125,776	339,016	62.9
Humanitarian	45,963	49,822	95,785	48.0
Total	695,480	312,919	1,008,399	69.0

The proportion of linked records may look low however analysis conducted in Section 2.4 indicates that these proportions are similar to the 2011 ACMID.

Year of arrival in Australia

Graph 2.12 shows the proportion of linked in-scope records for each reference period, by year of arrival. While the proportion of linked records is always higher than the proportion of unlinked records for each year of arrival (i.e. above 50%), the proportion of linked records grow slowly until 2006. From 2007 onwards, the percentage of linked records slowly decreases as the year of arrival grows more recent. This suggests that for more recent years the less time migrants have been in Australia, the less likely they are to appear in both datasets.

2.12 Proportion of linked in-scope records, By Year of arrival and reference period



Source: 2009/10 Personal Income Tax and Migrant Integrated Dataset
2010/11 Personal Income Tax and Migrant Integrated Dataset

In the 2005/06 financial year, the number of migrants on a Skilled visa increased by about 20,000 to around 100,000 and has remained at this level for several years (DIBP, 2014). Given that migrants on a Skilled visa are highly likely to be earning an income, this could account for the largest proportion of linked records around this time, however, without further investigation, this cannot be stated definitively.

Some reasons for why earlier arrival years have a lower proportion of linked records could be because some of the following reasons:

- Out of date address;
- Name change due to marriage or divorce; or
- May no longer be in the labour force or earning a declarable, taxable income.

Any of these attributes would make record linking more difficult.

2.3.4 Match rate estimates

Records in the SDB and PIT files are said to 'match' when they correspond to the same person. The match rate is the proportion of all matches that were correctly linked, and is calculated as:

$$\text{Match rate} = \frac{\text{Number of records correctly linked}}{\text{Number of matches}}.$$

It is not possible to know precisely how many of the links made between in-scope SDB records and the PIT were true. However, given that this feasibility study utilised Gold Standard linking and used a high cut-off to identify a match, it is reasonable to assume that a very high proportion of the links were correct. Therefore, for the purposes of the match rate analysis, it was assumed that all the in-scope links were correct.

There was also no way to determine the total number of matches which should be present between the in-scope SDB records and the PIT. As explained earlier, the scope limitations for the SDB meant that it was not possible to remove those persons whose income is below the tax free threshold, those who did not have tax withheld, or those who had not yet completed a PIT return. A number of secondary data sources were therefore used to estimate the likely number of matches that might be expected.

Adopting those assumptions and estimates, match rate estimates have been calculated by taking the number of SDB records that linked to a PIT record and dividing by the expected number of matches derived from the various other datasets (such as the 2011 ACMID). As a formula, this can be expressed as:

$$\text{Expected match rate} = \frac{\text{Number of SDB records correctly linked}}{\text{Expected number of SDB matches}}.$$

This section does not specifically give an indication of over- or under-representation of subpopulations. For more information on this, see Section 2.3.5.

2.3.4.1 Australian population aged 15–64 that submitted a tax return

The first consideration was the proportion of the population aged 15 to 64 years who submitted a tax return. The Estimated Resident Population (ERP) was used to calculate the estimated population (ABS, 2010).

2.13 Australian population who submitted a tax return, aged 15 to 64 years[(a)]

	2009/10	*2010/11*
Estimated total population	14,772,528	14,864,930
Number of PIT records	11,251,615	11,484,672
Proportion of population who submitted a tax return	76.2%	77.3%

(a) Age at the beginning of the relevant reference period (01/07/2009 for 2009/10 and 01/07/2010 for 2010/11) calculated from the PIT date of birth variable.

Using the proportion of the population who submitted a tax return, the expected number of matches can be calculated as:

$$\text{2009/10 Expected number of matches} = 891,290 \times 76.2\% = 679,163$$
$$\text{2010/11 Expected number of matches} = 1,008,399 \times 77.3\% = 779,492$$

Therefore, the match rate is calculated as:

$$\text{2009/10 Match rate} = \frac{613,584}{679,163} = 90.3\%$$
$$\text{2010/11 Match rate} = \frac{695,480}{779,492} = 89.3\% .$$

The match rate is estimated to be around 90% for both reference periods. However, there is no way to identify recent migrants using the Australian Demographic Statistics publication and the characteristics of recent migrants may differ substantially from the whole population. A better approximation of the expected number of links is needed.

2.3.4.2 Recent migrants aged 15–64 eligible to pay tax in the reference period using the Australian Census of Population and Housing, 2011

Comparing the proportion of SDB records linked to a PIT record against similar figures extracted from the *Australian Census of Population and Housing, 2011* (ABS, 2012) allows for a better estimate of the match rate as it provides an estimate of the recent migrant population.

People have been deemed eligible to pay tax if they reported a non-negative income on the 2011 Census.

2.14 2011 Census recent migrant records by eligibility to pay tax

	No. of records	*% of records*
Eligible to pay tax	1,197,942	77.5
Ineligible to pay tax[(a)]	307,064	19.9
Not stated	41,050	2.7
Total	1,546,056	100

(a) Includes 'Negative income' and 'Nil income'.

Using the proportion of the population considered eligible to pay tax, the expected number of matches can be calculated as:

$$\text{2010/11 Expected number of matches} = 1,008,399 \times 77.5\% = 781,509 .$$

Therefore, the match rate is calculated as:

$$\text{Match rate} = \frac{695,480}{781,509} = 89.0\% .$$

The match rate is again estimated to be almost 90%. However, as the category 'Eligible to pay tax' includes low income earners who may not have been required to submit a tax return and may not include those not required to pay tax but still receiving Centrelink payments, the match rate could be higher.

2.3.4.3 Recent Migrants aged 15–64 eligible to pay tax in the reference period using the Australian Census and Migrants Integrated Dataset (ACMID), 2011

Comparing the proportion of SDB records linked to a PIT record against similar figures extracted from the *Australian Census and Migrants Integrated Dataset (ACMID), 2011* (ABS, 2014a) allows for another estimate of the match rate to be calculated.

In the ACMID, recent migrants are defined as people born overseas who arrived in Australia between 1 January 2000 and 9 August 2011. This population only includes migrants on Skilled, Family and Humanitarian visas.

People on the ACMID have been deemed eligible to pay tax if they reported a non-negative income.

2.15 ACMID records by eligibility to pay tax, aged 15 to 64 years

	No. of records	*% of records*
Eligible to pay tax	776,223	80.7
Ineligible to pay tax[(a)]	161,429	16.8
Not stated	24,438	2.5
Total	962,089	100.0

(a) Includes 'Negative income' and 'Nil income'.

Using the proportion of the population considered eligible to pay tax, the expected number of matches can be calculated as:

$$\text{2010/11 Expected number of matches} = 1,008,399 \times 80.7\% = 813,778 .$$

Therefore, the match rate is calculated as:

$$\text{Match rate} = \frac{695,480}{813,778} = 85.5\% .$$

The match rate is estimated to be 85.5%. However, as the category 'Eligible to pay tax' includes low income earners who may not have been required to submit a tax return and may not include those not required to pay tax but still receiving Centrelink payments, the match rate could be higher.

2.3.5 *Under- or over-representation of sub-groups*

Various migrant characteristics from the SDB were compared with the PITMID file in order to understand differences in distribution of various subpopulations in the two files. In particular, the relative frequencies of subpopulations were compared. The analysis indicated that no major subpopulations were missed in the linking process when examining both the whole linked file and the records deemed as in-scope (see Appendix C). However, some groups appear to be more difficult to link than others, resulting in a small degree of under-representation on the linked file. It was found that rates of under-representation were highest for migrants in the following groups:

- Migrants on Family and Humanitarian visas;
- Migrants born in North Africa and the Middle East; and
- Migrants whose latest visa application was submitted offshore.

Some factors that may have influenced this under-representation for the above groups are the quality of the SDB records in terms of name and address, the rate of labour force participation and hence potential to earn a taxable income, and visa conditions that may preclude earning an income.

In comparison, various subpopulations have been over-represented in the linked file, both for the original SDB file and the in-scope records. It was found that rates of over-representation were highest for migrants in the following groups:

- Migrants on Skilled visas;
- Migrants aged 25–44 years;
- Migrants born in North West Europe; and
- Migrants whose latest visa application was submitted onshore.

The largest recorded values of over- and under-representation appear when looking at the location where the latest visa application was submitted. The fact that applications processed onshore are over-represented by almost 10 percentage points on the whole file suggest that these migrants have been in Australia long enough to establish

themselves, probably as temporary migrants for a period of time prior to gaining permanent residency and are thus more likely to be in the labour force or earning an income through other sources.

Migrants aged 25 to 44 years can reasonably be considered to be primarily working age, so are more likely to be earning a taxable income. Some factors that may have influenced this over-representation are migrants on a Skilled visa have a higher labour force participation due to filling gaps in the labour market or having skills in more high demand occupations. Skilled migrants may also be sponsored or have employment opportunities arranged prior to arrival and hence be more likely to earn an income. Also, North West European migrants may be more proficient in English and so be more likely to be established in the labour market, given that English proficiency is a condition in many Skilled visa applications. It should be noted, however, that some of these differences may simply reflect some cohorts having proportionally more taxpayers.

Given that the PIT file is a specific subset of the Australian population, it is to be expected that there will be noticeable differences in the characteristics of the original SDB file and the PITMID file. For this reason, analysis in Section 2.3.4 was restricted by Age, Visa stream, permanent residence start date and departure date. Proportions from the in-scope file could perhaps provide a better representation of over- and under-representation, however more analysis would be needed to state this definitively and is beyond the scope of this project.

Detailed distributions of subpopulations are presented in the tables below, which compare the relative frequencies of migrant characteristics for the whole SDB file, the in-scope SDB file, the whole linked file and the in-scope linked file.

Further investigation into the effect of any over- or under-representation may be needed and a weighting or calibration process may be required to eliminate the potential bias within the linked data.

2.16 Relative frequencies (%) in each Visa category, for the SDB and PITMID file

			In-scope records			
	Whole file		*2009/10*		*2010/11*	
Visa stream	*SDB*	*PITMID*	*SDB*	*PITMID*	*SDB*	*PITMID*
			%			
Family	32.4	28.7	33.6	30.9	33.6	30.7
Humanitarian	8.1	5.2	9.3	6.7	9.5	6.6
Skilled	55.0	60.8	57.0	62.4	56.9	62.7
Permanent – Other	0.2	0.2	–	–	–	–
Temporary	4.1	5.0	–	–	–	–

2.17 Relative frequencies (%) in each Age group, for the SDB and PITMID file

			In-scope records			
	Whole file		2009/10		2010/11	
Age group	SDB	PITMID	SDB	PITMID	SDB	PITMID
			%			
Under 15 years	6.7	0.3	–	–	–	–
15–17 years	3.7	2.0	4.6	2.5	4.7	1.6
18–24 years	15.2	15.5	11.1	9.7	10.7	9.0
25–34 years	36.6	42.9	38.0	40.2	36.7	39.4
35–44 years	23.0	26.6	29.8	32.3	30.5	33.4
45–54 years	9.2	9.6	12.6	12.5	13.2	13.6
55–64 years	3.3	2.2	3.9	2.8	4.2	3.0
65–84 years	2.1	0.8	–	–	–	–
85 years and over	0.1	0.0	–	–	–	–

2.18 Relative frequencies (%) in each Region of birth, for the SDB and PITMID file

			In-scope records			
	Whole file		2009/10		2010/11	
Region of birth	SDB	PITMID	SDB	PITMID	SDB	PITMID
			%			
Oceania and Antarctica	1.9	1.9	2.1	2.1	2.1	2.1
North West Europe	18.2	21.0	19.1	22.2	18.7	22.0
Southern and Eastern Europe	2.9	2.9	3.1	3.1	3.1	3.1
North Africa & Middle East	7.4	5.0	8.5	6.1	8.3	5.9
South East Asia	16.5	15.0	17.3	15.8	17.2	15.7
North East Asia	18.5	16.8	17.5	16.2	17.3	16.1
Southern and Central Asia	20.0	22.0	17.4	18.5	18.1	19.1
Americas	4.5	4.9	4.2	4.4	4.2	4.5
Sub Saharan Africa	8.7	9.2	9.0	9.9	9.4	10.1

2.19 Relative frequencies (%) in each visa application location, for the SDB and PITMID file

			In-scope records			
	Whole file		2009/10		2010/11	
Location of latest visa application	SDB	PITMID	SDB	PITMID	SDB	PITMID
			%			
Onshore	36.4	46.1	29.4	32.2	30.5	34.0
Offshore	63.6	53.9	70.7	67.8	69.5	66.0

2.4 Analysis of the linked dataset

In addition to measuring the match rate, an analysis was conducted on other variables in the linked dataset to see whether the data matched other data sources that contained migrant income information.

Despite the fact that the ACMID and ABS' Characteristics of Recent Migrants (CORMS) datasets have different populations, it is expected that the majority of people reporting an income would have submitted a tax return and would have had a corresponding PIT record. For this reason, it is possible to compare data from the PITMID against these sources.

As shown in the sections below, the PITMID linked dataset provided similar proportions with respect to total income and sources of income as several other migrant datasets.

2.4.1 Australian Census and Migrants Integrated Dataset (ACMID), 2011

The following table shows the total annual personal income of recent migrants aged 15–64 years from the ACMID (ABS, 2014a) and the Gross income variable for linked in-scope records from the PITMID.

2.20 Total annual personal income

		PITMID	
	ACMID	*2009/10*	*2010/11*
		%	
Negative income	1.2	1.3	1.2
Eligible to pay tax	98.9	98.7	98.8
$1–$10,399	11.0	11.9	10.5
$10,400–$15,599	9.5	7.2	6.4
$15,600–$20,799	6.1	6.5	6.5
$20,800–$31,199	11.3	11.8	11.2
$31,200–$41,599	13.5	13.4	12.7
$41,600–$51,999	11.2	11.9	11.8
$52,000–$64,999	10.5	11.0	11.2
$65,000–$77,999	7.9	7.9	8.4
$78,000–$103,999	8.8	8.9	10.1
$104,000 or more	9.1	8.4	10.1
Total	100.0	100.0	100.0

The similar distributions in all of the income categories show that no major subpopulations were missed in the linking process. Despite the fact that the ACMID has a different population, it is expected that the majority of people reporting an income would have submitted a tax return and would have had a corresponding PIT record.

Given that wages and salary are the main source of income for the majority of Australian residents, another analysis compared the proportion of in-scope SDB records that linked to a PIT record with the labour force participation rate from the ACMID. It should be noted that there are many different sources of income and this method is not exact, though it does give a good picture.

Using the figures from table 2.11, in the 2010/11 reference period 76.1% of migrants who held a Skilled visa linked to a PIT record, 62.9% of Family visas and 48% of Humanitarian visas. These figures were compared with the Labour force participation rates from migrants aged 15–64 on the ACMID who had arrived in Australia between 1 January 2000 and Census night.

2.21 Labour force participation rate, 2011 ACMID

	Visa stream			
	Skilled	*Family*	*Humanitarian*	*Total*
	(%)			
In the labour force	81.0	65.6	42.3	71.9
Not in the labour force	18.5	33.4	55.3	27.2
Not stated	0.5	1.0	2.4	0.85
Total	100.0	100.0	100.0	100.0

The figures for the labour force participation rate are very similar to the linkage rates. The larger difference concerning Humanitarian migrants (48% compared with 42%) could be explained by a larger proportion of Humanitarian migrants being dependent on Government pensions and benefits and being less likely to be in the labour force.

2.4.2 Characteristics of Recent Migrants Survey (CORMS), November 2010

The following table shows the main source of household income for migrants aged 15 to 64 years from the 2010 Characteristics of Recent Migrants Survey (ABS, 2011) and the two linked Client data files.

Main source of income is defined on both the CORMS and the PITMID as the income source from which the migrant received the most money.

2.22 Main source of income

		PITMID	
	CORMS 2010	*2009/10*	*2010/11*
	%		
Wages and Salaries	82.9	86.0	85.4
Other source	17.1	14.0	14.6

The figures between the two datasets are still quite similar, differing by no more than 3.2 percentage points, suggesting that the data aligns with what is already known about the population. However, it must be noted that income variables in CORMS are recorded as household income, while the Client data file relates to personal income.

2.5 Considerations for future iterations

Though the analysis in Section 2.3.4 indicated that the linkage was successful, a number of suggested changes to the process were identified. These changes could lead to an improvement in the quality of the linking in future iterations of the project.

2.5.1 TRIPS extract arrival and departure information

This feasibility study utilised the last movement direction (arrival or departure) that was provided for the ACMID project. This variable was used to exclude people whose last movement was a departure prior to the beginning of the reference period in the analysis conducted in Section 2.3.4.

The extract only contained records up to Census night (9 August 2011). This differed from the scope of the subsequent SDB extract for the Migrant PIT DI project, which included migrants up to 6 March 2013.

Future iterations of this project would benefit if TRIPS arrival and departure data containing more arrivals and departures than just the last movement direction for the reference period were obtained for the SDB extract. This data could then be used to exclude out of scope records prior to linking. This data would assist in the analysis of people who departed Australia temporarily prior to or during the reference period only to return again during the same period. This could indicate less time spent earning an income and may therefore more accurately reflect their eligibility to pay tax. However, this information may be too complex to extract.

2.5.2 Change scope of PAYG and Client data extracts

The ATO were asked to only supply records from the PAYG file which also had a record on the Client data file.

Future iterations of this project would benefit if the scope of these extracts could include any person for whom there was at least one employer-submitted wage and salary record for the reference period, regardless of whether they are present on the Client data file. This adjustment would improve the ability to assess the success of the linkage and provide information about low income earners.

2.5.3 Add name change sequence number in TRIPS extract

Names on the SDB extract are correct to the date of the latest permanent visa application. The case of women who change their surname after marrying or

divorcing presents a significant non-alignment issue. This feasibility study utilised the TRIPS Name history extract to overcome this non-alignment issue. This variable was used to populate missing or erroneous fields with legitimate values.

Future iterations of this project would benefit if a 'Name history sequence number' was present on future alias extracts. The availability of a name effective date, such as the one provided on the address file extract would also aid in determining the alias at the time of the reference period and could lead to more accurate linkage.

2.6 Evaluation summary

In summary, there were a number of valid reasons as to why a record on the SDB would not have a corresponding PIT record. As such, it was not expected that the linkage rate for this feasibility study would reach 100 per cent for the entire population. These reasons were outlined in Section 2.3.2.

At the end of the linkage process, 951,234 SDB records could be linked to the PIT file using variables such as name, address, sex and date of birth. This equated to 54% of the original SDB file.

Given that the two datasets are not subsets of each other, estimating the success of the linkage was difficult as not all of the out-of-scope records could be removed prior to the linking process.

However, limiting the available records by characteristics including age and visa stream to try to identify those most likely to submit a tax return showed that the proportion of SDB records that linked to a PIT record increased to around 70%. After identifying the same population in datasets such as the *Australian Census of Population and Housing, 2011* and the *Australian Census and Migrants Integrated Dataset (ACMID), 2011*, a similar proportion of the population were deemed "eligible to pay tax" and indicated that the true linkage rate could have been as high as 90%.

There are also several factors which may contribute to possible errors on the SDB and PIT. For some migrants, English proficiency may be a barrier to providing complete and correct information. Errors may also arise when a proxy provides information on behalf of the respondent.

Differences in distributions of subpopulations mean 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, as was done in ACMID.

While various adjustments could be made to future iterations of this project which could improve the accuracy of the links, the method used in this feasibility study has proved sufficient for the linkage methodology to be considered feasible.

3. POSSIBILITIES OF USING PIT DATA FOR OTHER LINKAGE PROJECTS

3.1 Gold standard linkage

In terms of linkage projects, the PIT data would be of great value for gold standard probabilistic linkage, where name and address details are used as linking variables. Name information may require some standardisation (as discussed in Section 2.2.2) however the variable has a high rate of completeness and is of high quality.

Address information is of particularly high quality. Addresses on the Name and Address register were geocoded to ABS standard geography categories of Meshblock, SA1 and SA2 (ABS, 2013a). Geocoding is reliant on full and accurate information to be coded correctly. Over 97% of the PIT records that linked to an SDB record for each reference period were successfully coded to a corresponding Meshblock and up to 99% were coded to an SA2 code, indicating a very high level of completeness and quality in the address information (see table 3.1).

3.1 Completeness rates for ABS geocoded variables for SDB linked records

	2009/10		*2010/11*	
	Client data	*PAYG*	*Client data*	*PAYG*
	% complete			
Meshblock	97.7	97.3	98.1	97.7
SA1	97.9	97.5	98.3	97.8
SA2	98.5	98.2	99.0	98.5

These values are slightly higher than the geocoding rates for the whole PIT file, which were 95.2% coded to a Meshblock and 98.1% coded to an SA2.

3.2 Bronze standard linkage

The PIT data may not be of great value in bronze standard linking projects. The register itself contains only basic information such as first name, any other given names, family name, address information, sex and date of birth. Spousal information is later appended to the file by the ATO via a match on the Scrambled Tax File Number (STFN).

There are a limited number of identifying variables on the Name and address register that could conceivably be used as linking variables if name and address are not utilised. Variables such as sex and date of birth would not be enough to definitively link records without duplicates and would result in a high level of multiple links.

3.3 One-to-one linkage

Although the PIT data contains a unique record ID of Tax File Number (TFN), specific legislation forbids the ATO from disclosing this record identifier.

Alternatively, having unique IDs from other datasets on the PIT dataset would be one solution, however this is highly unlikely.

Hence, probabilistic or deterministic linking would be required for any future linkage project.

3.4 Overall linkage ability

3.4.1 Timing

The ATO provides unit record data to the ABS approximately 18 months after the end of the financial year. This extract includes tax returns processed up to 16 months after the end of the financial year. Any returns processed after this date are not included in the extract.

As mentioned in Section 2.1.2, any name information is correct up to the date of extraction. This can pose issues in cases where names may change, such as women marrying and taking their husband's surname.

Similarly, it was also mentioned that any address information is correct up to the date of extraction. This too can pose issues due to the time it takes for the ABS to receive data from the ATO as respondents could change address during this period. The Migrant PIT DI project was able to mitigate this issue by obtaining both the Name history and Address history extracts from the SDB so that name and address changes could be taken into account.

These same potential issues exist with name and address changes which could be a factor in any other linkage project and a similar strategy would need to be developed to mitigate this issue.

3.4.2 Quality of other datasets

Given the unique nature of the PIT data items, any dataset being integrated with the PIT data will need to have detailed personal information such as those on the PIT dataset. For the Migrant PIT DI project, there were so few linking variables to both files that highly identifying information such as name, date of birth and address were essential to obtain high quality links.

4. POSSIBLE NEW STATISTICS ON RECENT MIGRANTS

4.1 Some areas of interest for research

There are a number of research areas of interest that the integrated data from the Migrant PIT DI project may address. These research questions are briefly discussed in this section and some ideas for analytical work to inform the questions are presented.

References to migrants are in the context of those who linked to an available PIT record indicating that they have submitted a tax return or earned a wage or salary in the financial year. The data has not been calibrated to reflect the total population of migrants who submitted a tax return and thus may not be representative of the whole of that population.

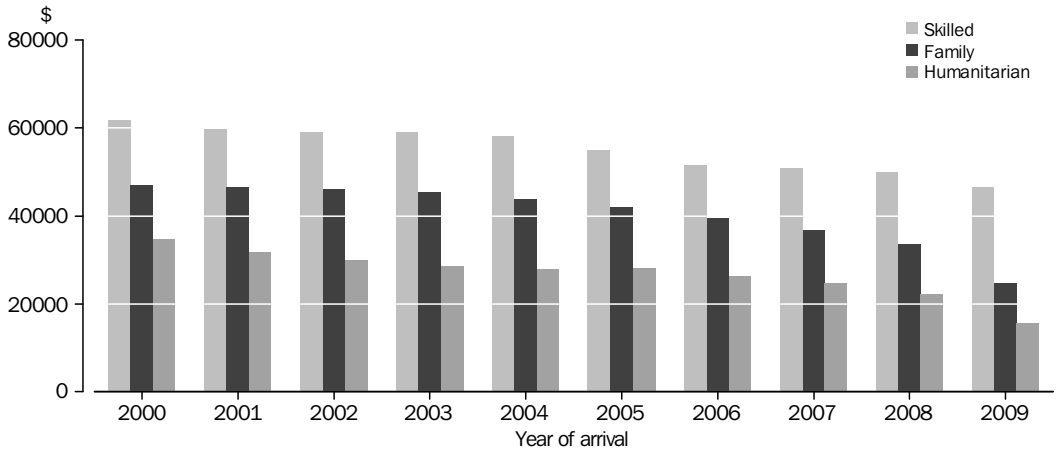
The linked datasets will be referred to as the 2009/10 Personal Income Tax and Migrants Integrated Dataset (PITMID) for the 2009/10 financial year and the 2010/11 PITMID for the 2010/11 financial year.

4.1.1 Economic outcomes

By linking the SDB records to PIT records, economic outcomes of recent migrants, namely income from Wages and Salary, can be cross-classified by migration related variables such as visa stream, country of birth and length of time since arrival in Australia. Detailed income data from all sources would then be available on different cohorts of migrants who enter Australia under different policy settings. This information would be of great benefit for evidence-based policy and research on migrants and their settlement outcomes. Over time, access to annual series and longitudinal data would allow migrant economic outcomes to be measured. Spatially enabled data would allow for regional analyses. However this sort of analysis would be subject to the migrant population in that area being of sufficient size for the data to be made available.

Graph 4.1 below shows the average annual income for the 2009/10 PITMID by visa stream. It can be clearly seen that in the 2009/10 financial year, Skilled visa holders earn the highest average incomes ($50,000 to $60,000), Family visa holders earned on average between $20,000 and $50,000 and Humanitarian visa holders earn the lowest average incomes, from $15,000 to $35,000.

4.1 Linked migrant taxpayer records, 2009/10 – Average annual income from Wages and Salaries, By Year of arrival and Visa stream

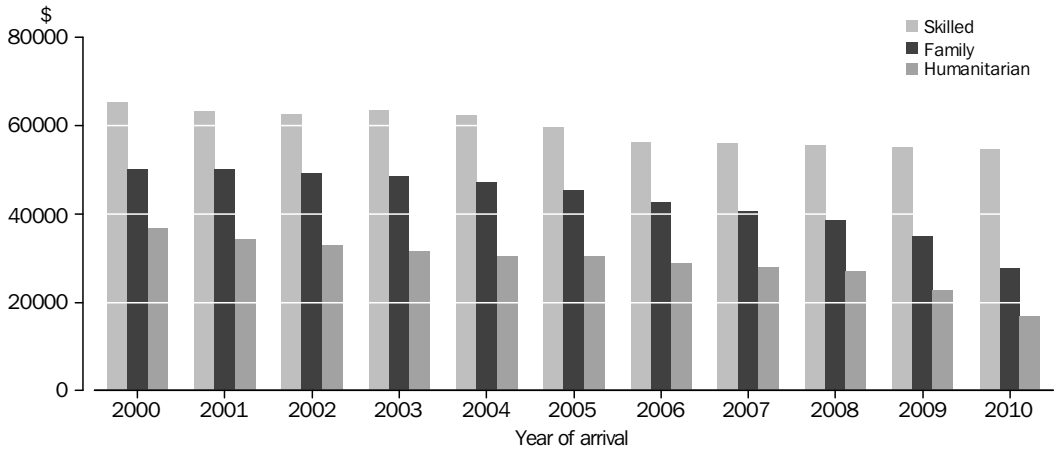


Note: These estimates are uncalibrated and are considered experimental.

Source: 2009/10 Personal Income Tax and Migrant Integrated Dataset

In graph 4.2, the 2010/11 PITMID average annual incomes of Skilled visa holders rose to between $55,000 and $65,000. Family visa holders earnings also rose slightly in the 2010/11 financial year to average between $30,000 and $50,000. The income levels for Humanitarian visa holders remained within the same range, however the average income from wages and salary across all arrival years rose from $27,000 in 2009/10 to $29,000 in 2010/11.

4.2 Linked migrant taxpayer records, 2010/11 – Average annual income from Wages and Salaries, By Year of arrival and Visa stream



Note: These estimates are uncalibrated and are considered experimental.

Source: 2010/11 Personal Income Tax and Migrant Integrated Dataset

Consistently, the average income from wages in salary decreases as the year of arrival in Australia grows more recent.

4.1.2 Occupation outcomes

One area of interest for analysis is the occupation of the employee compared with the industry of the employer. Occupation is recorded on the ATO Client data file and is coded according to the *Australian and New Zealand Standard Classification of Occupations* (ABS, 2013b). Industry is recorded on the PAYG file and is coded according to the *Australian and New Zealand Standard Industrial Classification* (ABS, 2013c).

There is only one opportunity for Occupation to be recorded on the ATO Client data file and the ATO PAYG file contains multiple records with different industry information for migrants with multiple employers. This sort of information could either be analysed for migrants with only one record on the PAYG file or their main job on the PAYG file, i.e. the job that earned them the most money in the financial year.

In the future, a linked employee to employer dataset would enable the occupation of the employee to be connected with the industry of the employer over time. This would be useful given that many visa classes are underpinned by labour market needs.

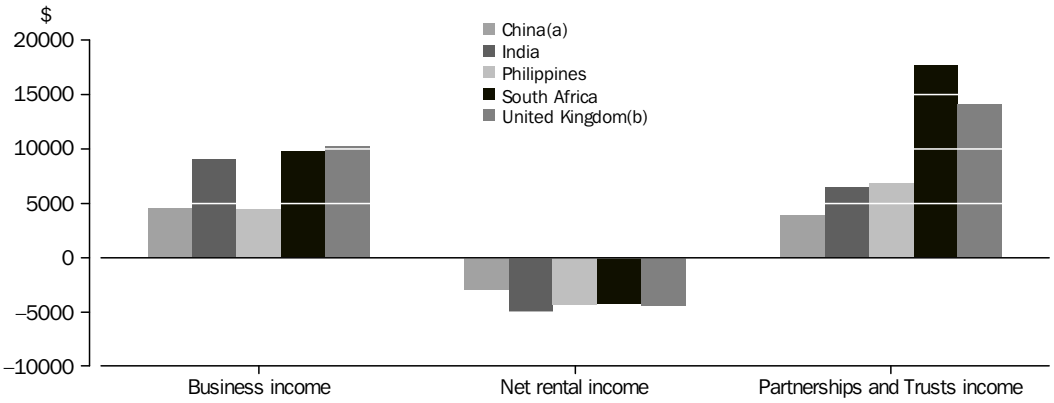
While comparisons of field of occupation and the qualifications held by the migrant on arrival are of interest, there is no information on the level of qualification held by the migrant on the SDB to enable this sort of analysis.

4.1.3 Business income

There is also interest in the amount of income generated from entrepreneurial ventures such as businesses, partnerships, trusts and property rental. Analysis could reveal whether some migrant cohorts have a propensity towards entrepreneurial activities. This is most easily demonstrated by looking at those migrants who start their own business after settlement in Australia.

Graphs 4.3 and 4.4 show the average income from business enterprises for the top five countries of birth for both the 2009/10 and 2010/11 financial years.

4.3 Linked migrant taxpayer records, 2009/10 – Average annual income from business enterprises, By selected Country of birth and Type of business income

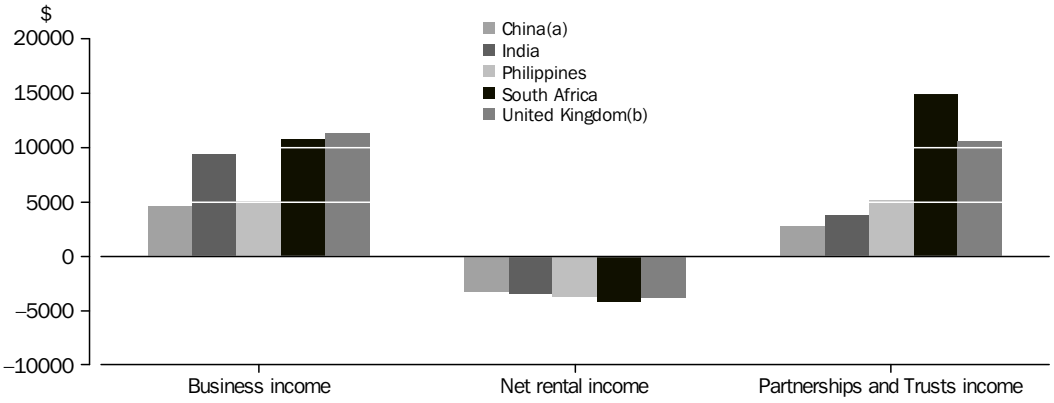


(a) Excludes SARs and Taiwan.
(b) Includes England, Wales, Northern Ireland, Scotland, the Channel Islands and the Isle of Man.

Note: These estimates are uncalibrated and are considered experimental.

Source: 2009/10 Personal Income Tax and Migrant Integrated Dataset

4.4 Linked migrant taxpayer records, 2010/11 – Average annual income from business enterprises, By selected Country of birth and Type of business income



(a) Excludes SARs and Taiwan.
(b) Includes England, Wales, Northern Ireland, Scotland, the Channel Islands and the Isle of Man.

Note: These estimates are uncalibrated and are considered experimental.

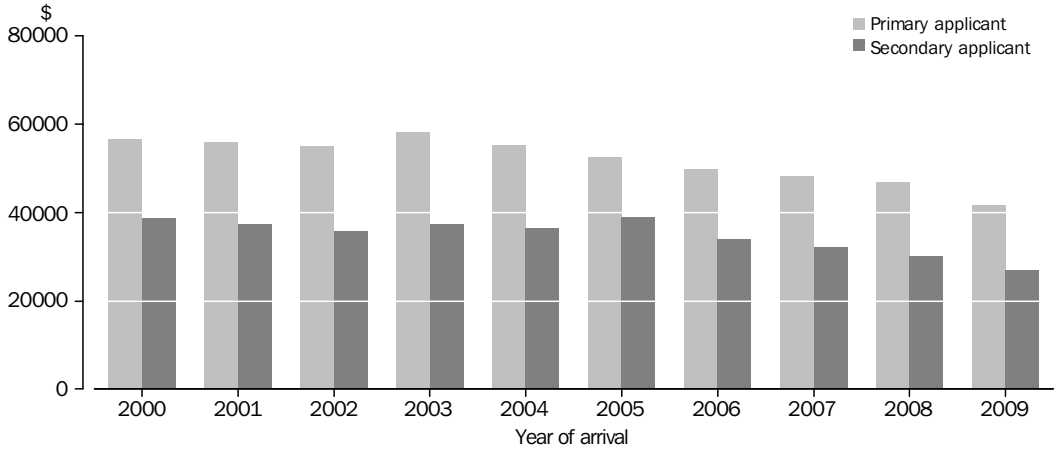
Source: 2010/11 Personal Income Tax and Migrant Integrated Dataset

4.1.4 Secondary applicant

There are two ways to analyse the income of secondary applicants. Firstly, the SDB *Primary (or main) Applicant Flag* can be used in conjunction with any of the PIT variables. This variable splits the records into Primary and Secondary applicants. The second option is to use *Spouse Taxable Income* from the ATO Client data file. This variable is only available for the 2009/10 reference period at the time of this study and does not contain the details of the sources of that total income. This variable may also be unavailable in future years.

The following graphs show the average annual total taxable income for linked migrant tax-payer records by applicant status using the primary applicant flag variable from the SDB for the 2009/10 and 2010/11 financial years. It is apparent that for both financial years that the income of the secondary applicant is consistently about two-thirds of the primary applicant.

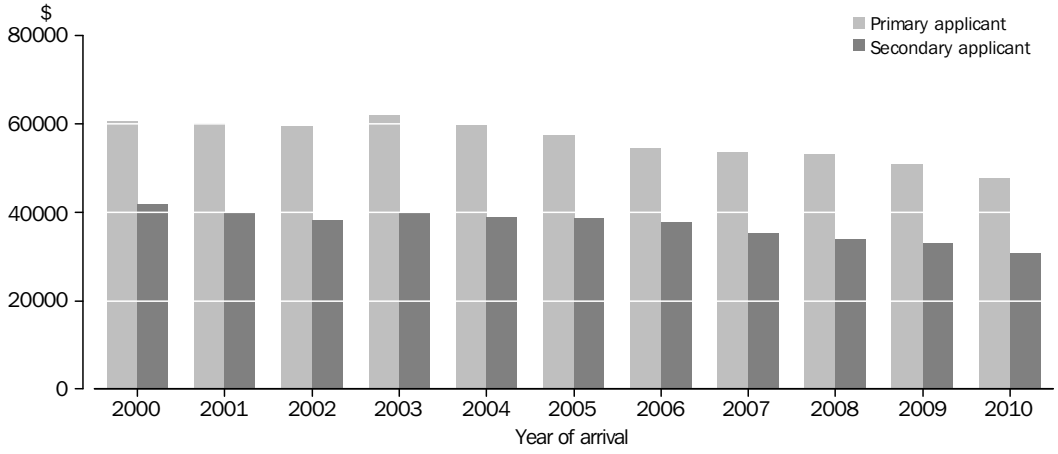
4.5 Linked migrant taxpayer records, 2009/10 – Average annual taxable total income, By Year of arrival and Applicant status



Note: These estimates are uncalibrated and are considered experimental.

Source: 2009/10 Personal Income Tax and Migrant Integrated Dataset

4.6 Linked migrant taxpayer records, 2010/11 – Average annual taxable total income, By Year of arrival and Applicant status



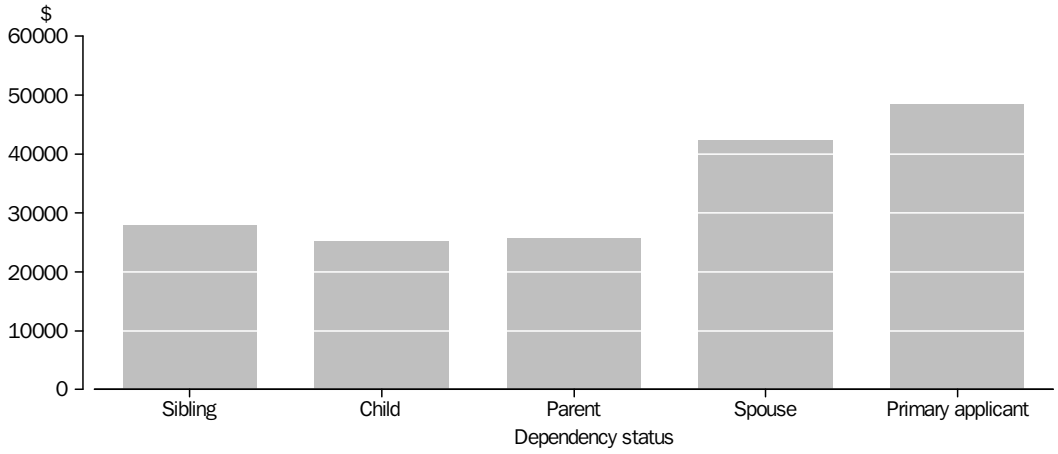
Note: These estimates are uncalibrated and are considered experimental.

Source: 2010/11 Personal Income Tax and Migrant Integrated Dataset

Also available on the SDB for secondary applicant analysis is the *Dependency* variable. This variable breaks up secondary applicants based on their relationship to the primary applicant. However, this variable is only available for offshore visa applicants. Consequently, the income variables for offshore dependent migrants cannot be compared with those with onshore status.

The following graphs show the average total taxable income for both primary applicants and secondary applicants by dependency status using the dependency variable on the SDB.

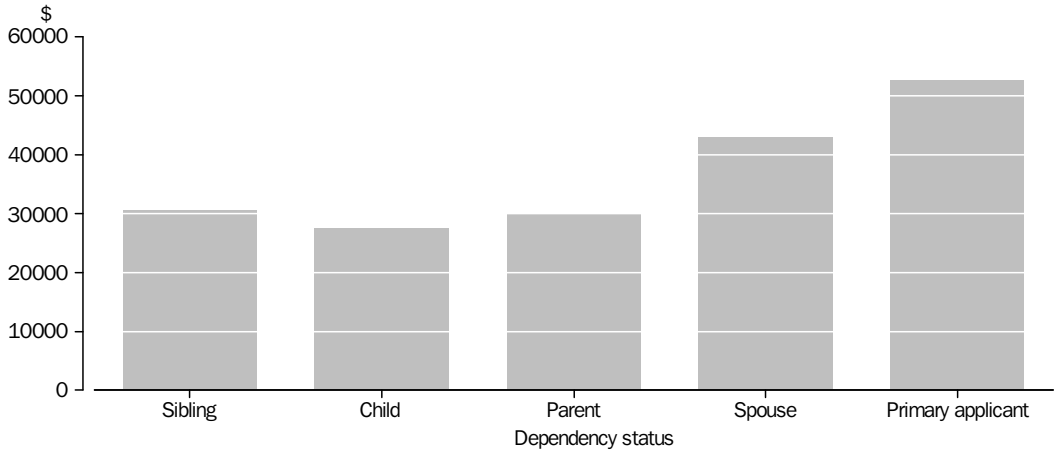
4.7 Linked migrant taxpayer records, 2009/10 – Average total taxable income for offshore migrants, By Dependency status and Primary applicant



Note: These estimates are uncalibrated and are considered experimental.

Source: 2009/10 Personal Income Tax and Migrant Integrated Dataset

4.8 Linked migrant taxpayer records, 2010/11 – Average total taxable income for offshore migrants, By Dependency status and Primary applicant



Note: These estimates are uncalibrated and are considered experimental.

Source: 2010/11 Personal Income Tax and Migrant Integrated Dataset

4.1.5 Spatially enabled data

In the longer term, provision of spatially enabled and longitudinal data would reveal the movement patterns of migrants over time provided that their first address location can be identified on the PIT data. However, as mentioned in Section 4.1.1, this sort of analysis would be subject to the size of the migrant population of interest.

Spatially enabled data would allow for analyses of regional migration programs to assess their effectiveness in building work forces in particular areas and generating regional population stability or growth, thus providing data to inform the national Long Term Migration Planning Framework.

Addresses on both the SDB and the PIT files are geocoded to Mesh Blocks as part of the linkage process and geographical variables on the analysis file are then provided at SA1 and SA2 level. Under current recommendations, those records classified as "high links" will not go through the linkage or geocoding process each successive year. This would result in these records not being updated for current address information each year. To overcome this, information could be requested each year from the SDB and PIT for all records including "high" linked records, and these could be geocoded. As the size of the file increases over time, the level of resources required to undertake this task may mean that this approach is not feasible. Other ways of updating this address information may become available in the future to offset this issue.

4.2 Other new statistics

In addition to the potential analyses outlined above, the PIT files contain a large number of other variables that could also be used to provide an insight into the income characteristics of recent migrants.

4.2.1 Foreign income and source(s)

Other possible interesting statistics could come from data items relating to foreign income sources. These include:

- Foreign employment income;
- Foreign rent;
- Foreign investment funds or life insurance policies; and
- Whether a migrant owns or has an interest in assets outside Australia with a total value of AUD$50,000 or more.

Some interesting analysis could be undertaken by Country of birth to see whether migrants from certain regions are more likely to display financial ties to other countries. It could also be useful to look at the number of foreign income sources by year of arrival to see if these financial ties lessen in relation to a longer duration of stay in Australia or whether they increase in relation to improvement in migrants' financial situations.

4.2.2 Health fund membership

There is potential to look at types of health fund membership in terms of Hospital cover, Extras cover or combined forms of cover. Analyses could examine whether there are patterns between individual levels of income and the type of health fund membership.

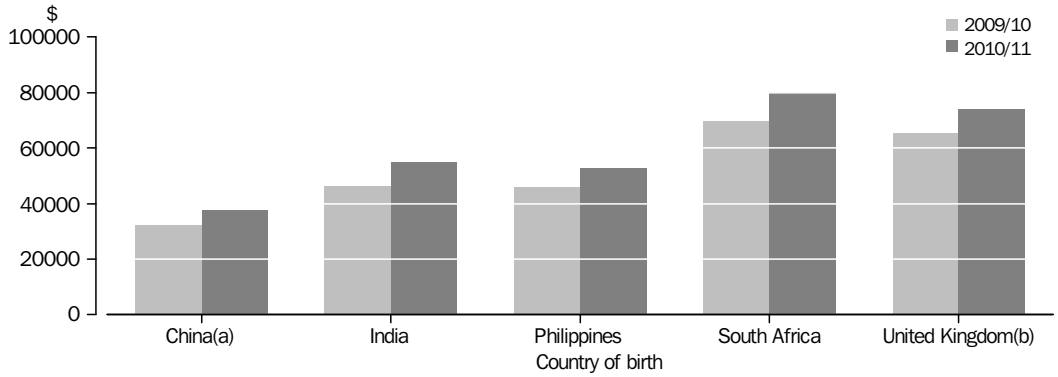
4.2.3 Longitudinal analysis

Potential longitudinal analyses could include income analysed by Country of birth or visa stream to identify which sub-populations have increased or decreased their income over time.

In addition, analyses of the spatial distribution of recent migrants in the context of their employment could be undertaken to assess the effectiveness of migrant programs such as those with a regional dimension. It would also be possible to determine whether migrants moved up the "occupation hierarchy" with increased periods of residency in Australia.

Graph 4.9 below shows the average total income for linked migrant tax-payer records for the top five countries of birth from both financial years. The graph presents only the records that were present in both files. Whilst there is an evident increase across the board for all of the countries from 2009/10 to 2010/11, India, South Africa and the United Kingdom had larger increases while China and the Philippines had smaller increases in income.

4.9 Linked migrant taxpayer records, 2009/10 and 2010/11 – Average total taxable income, By selected Country of birth



(a) Excludes SARs and Taiwan.
(b) Includes England, Wales, Northern Ireland, Scotland, the Channel Islands and the Isle of Man.

Note: These estimates are uncalibrated and are considered experimental.

Source: 2009/10 Personal Income Tax and Migrant Integrated Dataset
2010/11 Personal Income Tax and Migrant Integrated Dataset

REFERENCES

Australian Bureau of Statistics (2010) *Australian Demographic Statistics*, cat. no. 3101.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0 >

—— (2011) *Characteristics of Recent Migrants*, cat. no. 6250.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/6250.0 >

—— (2012) *Australian Census of Population and Housing, 2011*, ABS, Canberra.
< http://www.abs.gov.au/websitedbs/censushome.nsf/home/data >

—— (2013a) *Australian Statistical Geography Standard (ASGS): Volume 3 – Non ABS Structures*, cat. no. 1270.0.55.003, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.003 >

—— (2013b) *ANZSCO – Australian and New Zealand Standard Classification of Occupations, 2013, Version 1.2*, cat. no. 1220.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1220.0 >

—— (2013c) *Australian and New Zealand Standard Industrial Classification (ANZSIC), 2006, Revision 2.0*, cat. no. 1292.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1292.0 >

—— (2013d) *Estimates of Personal Income for Small Areas, Time Series, 2005-06 to 2010-11*, cat. no. 6524.0.55.002, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/6524.0.55.002 >

—— (2013e) *Wage and Salary Earner Statistics for Small Areas, Time Series, 2005-06 to 2010-11*, cat. no. 5673.0.55.003, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/5673.0.55.003 >

—— (2014a) *Microdata: Australian Census and Migrants Integrated Dataset, 2011,* cat. no. 3417.0.55.002, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/3417.0.55.001 >

—— (2014b) *Labour Force, Australia*, cat. no. 6202.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/6202.0 >

Department of Immigration and Border Protection (2014) *Historical Migration Statistics*, DIBP website, DIBP, Canberra.
< http://www.immi.gov.au/media/statistics/historical-migration-stats.htm >

Department of Immigration and Citizenship (2013) *Settlement Reporting Facility SRF Data Dictionary (External), Version 2*, DIAC, Canberra.
< http://www.immi.gov.au/living-in-australia/delivering-assistance/settlement-reporting-facility/pdf/ext-data-dictionary.pdf >

Department of Social Services (2014) *Social Security Payments – Residence Criteria*, DSS website, DSS, Canberra.
< http://www.dss.gov.au/about-the-department/international/policy/social-security-payments-residence-criteria >

Jaro, M.A. (1989) "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, 84(406), pp. 414–420.

Levenshtein, V.I. (1966) "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals", *Soviet Physics Doklady*, 10(8), pp. 707–710.

Richter, K.; Saher, G. and Campbell, P. (2013) "Assessing the Quality of Linking Migrants Settlement Records to 2011 Census Data", *Methodology Research Papers*, cat. no. 1351.0.55.043, Australian Bureau of Statistics, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.043 >

All URLs last viewed on 8 August 2014

APPENDICES

A. LINKING PASSES

A.1 Linking passes

	Deterministic																						*Prob.*	
Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23[g]	24[h]
Name information																								
Original first name (incl. multi-part strings)									E	E														
Clean first name (single string)																						E		
Clean first name (incl. multi-part strings)																							W85	W85
Standardised first name (single string)					E	E																		
Standardised first name (incl. multi-part strings)	E		E	E							E	W85	E	E			W85	W85		E[e]	W90[f]			
Standardised alias 2 first name (incl. multi-part strings)							E									E								
Standardised middle name (incl. multi-part strings)					[b]	[b]																		
Original last name									E	E														
Clean last name (incl. multi-part strings)	E		E	E	E	E					E	E	W85				W85			E[e]	E[f]	E		
Clean alias 2 last name (incl. multi-part strings)							E									E							W85	W85
All original name components in alphabetical order		E													E				W85					
All alias 2 name components in alphabetical order								E																
Personal characteristics																								
Sex	E	E	E	E	E	E	E	E	E	E	E	E	E	E[c]	E	E	E	E[c]	E	E	E	E	B	
Date of birth	E	E	E	E	E	E	E	E	E	E		E	E	E[d]	E	E	E	E[d]	E	E	E	E	B	
Year of birth											E													
Address information[a]																								
Street number (single string)											E	E	E	E	E	E	E	E	E					
Street number (incl. multi-part strings)	E	E			E		E	E	E															
Street name	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E				W85	W85
Mesh block			E			E				E														
SA2											E	E	E											
Postcode				E										E										
Spouse information																								
Original first name (incl. multi-part strings)																						E		
Clean last name (incl. multi-part strings)																						E		
Date of birth																						E		
Sex																						E		

E: Exact match.
B: Blocking variable.
W85: Winkler score of 0.85 required for a match.
W90: Winkler score of 0.90 required for a match.
(a) Where address information has been used for a pass, 3 passes were undertaken on SDB current, previous and forwarding information. For more information, see Section 2.2.6.
(b) Levenshtein distance was applied to the middle name variable. For more information, see Section 2.2.6.
(c) Sex was required to be female.
(d) Date of birth was required to not be an "administrative date". For more information, see Section 2.2.6.
(e) Required to have at least three name components.
(f) Required to have at least two name components.
(g) High acceptance threshold was used.
(h) Low acceptance threshold was used.

B. MISSING DATA RATES ON SDB AND PIT CANDIDATE LINKING VARIABLES

B.1 Missing data rates on SDB and PIT candidate linking variables

			PIT missing			
	SDB missing		2009/10 extract		2010/11 extract	
Variable	No. of records	(%)	No. of records	(%)	No. of records	(%)
Given names	38,687	1.9				
First given name[b]			381	0.0	442	0.0
Other given names[b]			4,609,354	37.1	4,771,148	37.5
Surname or family name	0	0.0	0	0.0	0	0.0
Date of birth	0	0.0	63	0.0	141	0.0
Sex[a]	58	0.0	0	0.0	0	0.0
Address line 1	762,358	18.3	39,434	0.3	39,889	0.3
Address line 2[c]	4,017,732	96.4	11,776,351	94.7	12,067,605	94.8
Address line 3 (Suburb)	757,260	18.2	38,337	0.3	38,845	0.3
State	20	0.0	19,444	0.2	19,600	0.2
Postcode	0	0.0	283	0.0	345	0.0
Country[d]			12,306,286	99.0	12,605,475	99.1
Spouse given name[b][e]			6,221,492	50.0	6,383,595	50.2
Spouse surname or family name[b][e]			6,219,893	50.0	6,381,530	50.2
Spouse date of birth[b][e]			6,443,271	51.8	6,574,615	51.7
Spouse sex[b][e]			6,447,762	51.9	6,559,697	51.6

(a) Missing responses for Sex were imputed based on Given names.
(b) Variable is only available on the PIT file.
(c) The high rate of missing values for this variable is to be expected and does not have any effect on the linkage as it is supplementary information.
(d) The high rate of missing values for this variable is to be expected as it indicates residential address is in Australia.
(e) The high rate of missing values for this variable is to be expected and may indicate the respondent is unmarried.

C. BREAKDOWN OF 'IN-SCOPE' RECORDS

In an effort to identify the population most likely to have submitted a tax return, the dataset was restricted by Age, Visa stream, Permanent residence start date and Permanent departure date. These four variables, in the context of the SDB, are analysed independently of each other in the sections below.

Age

After removal of the out-of-scope records, there were 1,614,632 records in the 2009/10 reference period and 1,633,819 records in the 2010/11 reference period aged 15 to 64 years.

C.1 Linked SDB records by age for 2009/10 reference period

		No. of records			*% of input file*	
Scope	*Age*[a]	*Input file*	*High links*	*All links*	*High links*	*All links*
In scope	15–17 years	64,999	7,925	10,294	12.2	15.8
	18–24 years	270,018	89,455	114,426	33.1	42.4
	25–34 years	649,372	295,594	353,746	45.5	54.5
	35–44 years	408,099	188,798	225,453	46.3	55.2
	45–54 years	163,940	68,992	82,698	42.1	50.4
	55–64 years	58,204	14,928	17,867	25.6	30.7
	Sub-total	1,614,632	665,692	804,484	41.2	49.8
Out of scope	Under 15 years	119,111	762	962	0.6	0.8
	65–84 years	37,979	5,881	6,809	15.5	17.9
	85 years and over	1,611	193	251	12.0	15.6
	Sub-total	158,701	6,836	8,022	4.3	5.1
Total		1,773,333	672,528	812,506	37.9	45.8

(a) Age at the beginning of the relevant reference period (01/07/2009 for 2009/10) calculated from the SDB date of birth variable.

C.2 Linked SDB records by age for 2010/11 reference period

		No. of records			*% of input file*	
Scope	*Age*[a]	*Input file*	*High links*	*All links*	*High links*	*All links*
In scope	15–17 years	66,695	9,174	11,605	13.8	17.4
	18–24 years	232,647	86,818	110,260	37.3	47.4
	25–34 years	651,485	322,032	381,650	49.4	58.6
	35–44 years	436,283	215,984	255,884	49.5	58.7
	45–54 years	182,699	83,367	99,059	45.6	54.2
	55–64 years	64,010	18,189	21,627	28.4	33.8
	Sub-total	1,633,819	735,564	880,085	45.0	53.9
Out of scope	Under 15 years	96,167	812	994	0.8	1.0
	65–84 years	41,322	6,921	7,872	16.7	19.1
	85 years and over	2,025	236	298	11.7	14.7
	Sub-total	139,514	7,969	9,164	5.7	6.6
Total		1,773,333	743,533	889,249	41.9	50.1

(a) Age at the beginning of the relevant reference period (01/07/2010 for 2010/11) calculated from the SDB date of birth variable.

Visa stream

Permanent visa stream information does not change from year to year. After removal of the out-of-scope records, there were 1,694,881 records holding a Skilled, Family or Humanitarian visa for both reference periods.

C.3 Linked SDB records by visa stream for 2009/10 reference period

		No. of records			*% of input file*	
Scope	*Visa stream*	*Input file*	*High links*	*All links*	*High links*	*All links*
In scope	Skilled	975,629	425,312	507,786	43.6	52.0
	Family	574,934	189,514	226,006	33.0	39.3
	Humanitarian	144,318	27,112	38,156	18.8	26.4
	Sub-total	1,694,881	641,938	771,948	37.9	45.5
Out of scope	Permanent – Other	3,372	793	1,361	23.5	40.4
	Provisional	72,823	28,699	37,876	39.4	52.0
	Not elsewhere classified	243	19	34	7.8	14.0
	Not applicable	2,014	1,079	1,287	53.6	63.9
	Sub-total	78,452	30,590	40,558	39.0	51.7
Total		1,773,333	672,528	812,506	37.9	45.8

C.4 Linked SDB records by visa stream for 2010/11 reference period

		No. of records			*% of input file*	
Scope	*Visa stream*	*Input file*	*High links*	*All links*	*High links*	*All links*
In scope	Skilled	975,629	462,238	547,028	47.4	56.1
	Family	574,934	213,751	251,702	37.2	43.8
	Humanitarian	144,318	32,214	44,392	22.3	30.8
	Sub-total	1,694,881	708,203	843,122	41.8	49.7
Out of scope	Permanent – Other	3,372	853	1,453	25.3	43.1
	Provisional	72,823	33,365	43,344	45.8	59.5
	Not elsewhere classified	243	18	34	7.4	14.0
	Not applicable	2,014	1,094	1,296	54.3	64.3
	Sub-total	78,452	35,330	46,127	45.0	58.8
Total		1,773,333	743,533	889,249	41.9	50.1

Permanent residence start date

After removal of the out-of-scope records, there were 1,105,392 records in the 2009/10 reference period and 1,269,739 records in the 2010/11 reference period whose permanent residence start date was prior to the reference period.

C.5 Linked SDB records by permanent residence start date for 2009/10 reference period

			No. of records			*% of input file*	
Scope	*Permanent residence start date*	*Location*	*Input file*	*High links*	*All links*	*High links*	*All links*
In scope	Before reference period	Onshore	306,309	152,963	190,091	49.9	62.1
		Offshore	799,083	311,951	382,078	39.0	47.8
	Sub-total		1,105,392	464,914	572,169	42.1	51.8
Out of scope	During reference period	Onshore	66,194	37,856	44,030	57.2	66.5
		Offshore	98,153	25,436	29,123	25.9	29.7
	After reference period	Onshore	272,849	130,347	149,242	47.8	54.7
		Offshore	230,745	13,975	17,942	6.1	7.8
	Sub-total		667,941	207,614	240,337	31.1	36.0
Total			1,773,333	672,528	812,506	37.9	45.8

C.6 Linked SDB records by permanent residence start date for 2010/11 reference period

			No. of records			*% of input file*	
Scope	*Permanent residence start date*	*Location*	*Input file*	*High links*	*All links*	*High links*	*All links*
In scope	Before reference period	Onshore	372,503	192,882	234,192	51.8	62.9
		Offshore	897,236	365,839	442,054	40.8	49.3
	Sub-total		1,269,739	558,721	676,246	44.0	53.3
Out of scope	During reference period	Onshore	86,149	53,562	61,194	62.2	71.0
		Offshore	78,352	19,563	22,747	25.0	29.0
	After reference period	Onshore	186,700	100,201	114,398	53.7	61.3
		Offshore	152,393	11,486	14,664	7.5	9.6
	Sub-total		503,594	184,812	213,003	36.7	42.3
Total			1,773,333	743,533	889,249	41.9	50.1

TRIPS departure information

After removal of the out-of-scope records, there were 1,414,659 records in the 2009/10 reference period and 1,627,222 records in the 2010/11 reference period whose last movement was not a permanent departure prior to the reference period.

C.7 Linked SDB records by departure information for 2009/10 reference period

		No. of records			*% of input file*	
Scope	*Departure information*	*Input file*	*High links*	*All links*	*High links*	*All links*
In scope	Departure during reference period	42,192	3,725	7,077	8.8	16.8
	Arrival during reference period	212,563	102,446	119,946	48.2	56.4
	No departure information	1,414,659	565,342	683,108	40.0	48.3
	Sub-total	1,669,414	671,513	810,131	40.2	48.5
Out of scope	Departure prior to beginning of reference period	103,919	1,015	2,375	1.0	2.3
Total		1,773,333	672,528	812,506	37.9	45.8

C.8 Linked SDB records by departure information for 2010/11 reference period

		No. of records			*% of input file*	
Scope	*Departure information*	*Input file*	*High links*	*All links*	*High links*	*All links*
In scope	Departure during reference period	91,632	15,302	22,541	16.7	24.6
	Arrival during reference period	504,707	281,714	331,764	55.8	65.7
	No departure information	1,030,883	444,197	530,324	43.1	51.4
	Sub-total	1,627,222	741,213	884,629	45.6	54.4
Out of scope	Departure prior to beginning of reference period	146,111	2,320	4,620	1.6	3.2
Total		1,773,333	743,533	889,249	41.9	50.1

Final in-scope records

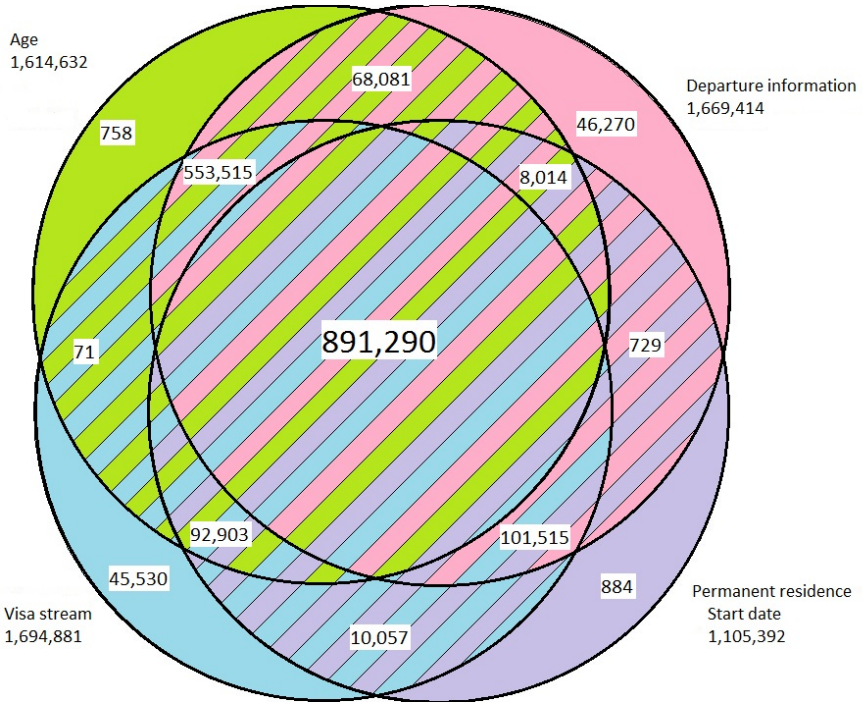
Records were considered to be in-scope if they met the in-scope criteria for all four of the variables outlined above. However, there were several variables that were in-scope for at least one variable, yet out of scope for others.

The diagrams below show the number of SDB records in relation to the degree to which they can be considered 'in-scope' with respect to the four variables outlined above. For example, 68,081 SDB records were considered 'in-scope' of both Age and Departure information but were 'out-of-scope' for both Visa stream and Arrival date.

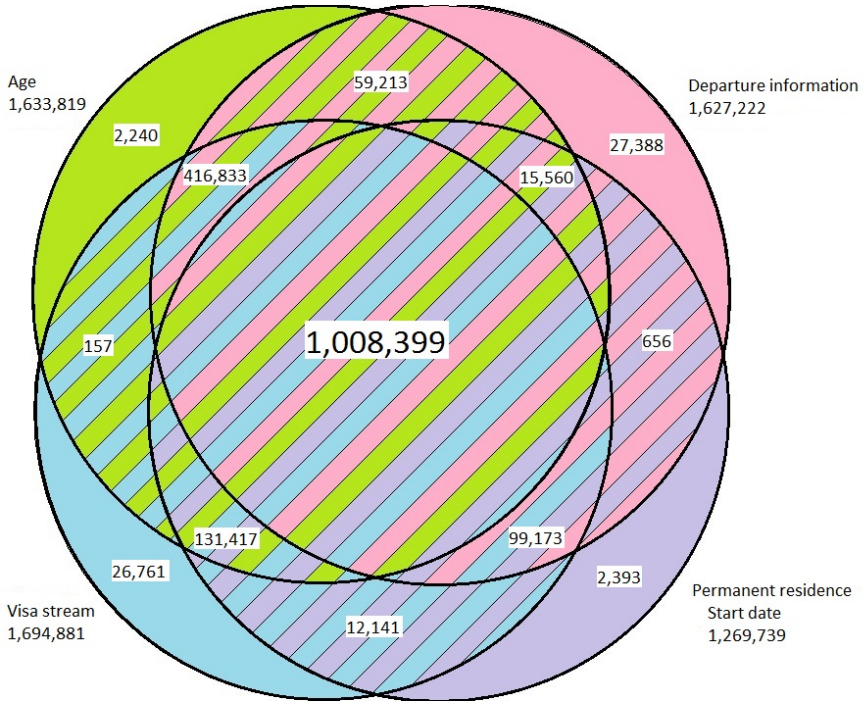
Consequently there were 891,290 person SDB records that met the in-scope criteria for all four of the variables for the 2009/10 period and 1,008,399 person SDB records in scope for the 2010/11 period. These figures are used to calculate the match rates in Section 2.3.4.

The final number of linked records, 613,584 for 2009/10 and 695,480 for 2010/11, along with the proportion of linked in-scope records, is shown in table 2.7.

C.9 In-scope SDB records for 2009/10 reference period



C.10 In-scope SDB records for 2010/11 period



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