1351.0.55.053

Australian Bureau of Statistics

Research Paper

Assessing the Suitability of Temporary Migrants Administrative Data for Data Integration

www.abs.gov.au

New Issue

Research Paper

Assessing the Suitability of Temporary Migrants Administrative Data for Data Integration

National Migrant Statistics Unit

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) WED 5 NOV 2014

ABS Catalogue no. 1351.0.55.053

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 Migrant Statistics Branch, on Adelaide (08) 8237 7402.

ASSESSING THE SUITABILITY OF TEMPORARY MIGRANTS ADMINISTRATIVE DATA FOR DATA INTEGRATION

National Migrants Statistics Unit
Australian Bureau of Statistics

EXECUTIVE SUMMARY

In 2014, the Australian Bureau of Statistics (ABS) National Migrants Statistics Unit (NMSU) conducted a study to examine the quality of the Department of Immigration and Border Protection (DIBP) Temporary Visa Holders (TVH) administrative data in terms of its suitability for data integration. For the purposes of this study, the administrative data on Temporary migrants is composed of data on International Students and Temporary work (skilled) (subclass 457) visa holders as at 31 July 2011.

The study investigates the quality of the administrative data of migrants on student and Temporary work (skilled) visas as well as assessing the suitability of the data for linking with the (ABS) 2011 Australian Census of Population and Housing and the Australian Taxation Office (ATO) Personal Income Tax (PIT) data. If this data were linked to the Census data file, a linked Australian Census and Temporary Migrants Integrated Dataset (ACTMID) could provide enhanced information on temporary migrant outcomes. If an integrated dataset could be produced of sufficient quality, then the dataset would provide new information on the characteristics of temporary migrants such as usual residence currently and one year ago, labour force status, educational qualifications obtained, income and housing. This data would be very useful for policy and planning as well as providing a rich source of data on this growing population group for academics and researchers to inform their studies.

The study also explores methodologies to improve future data integration studies and compares the findings with earlier quality studies that have assessed the quality of the integrated dataset created by linking the 2011 Census to the Department of Social Services (DSS) and Department of Immigration and Border Protection (DIBP) Settlement Database (SDB).

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.

The characteristics of the data items provided for the study were assessed for completeness and any anomalies. The feasibility of linking with the 2011 Census was assessed by an experimental linkage simulation tool. This tool simulates the probabilistic linking process and provides a diagnostic report enabling the feasibility of linking two datasets to be assessed prior to linkage.

This paper provides background to the Temporary Migrant Feasibility Study, a discussion of the quality of the TVH administrative data items, and a brief description of the simulated linking strategy, with detailed results included in the Appendix.

The results of the quality study and the linkage simulation show that linking the TVH data to the 2011 Census is likely to result in a link rate of about 70% but that there are a number of issues with the data. The most prevalent issue with the data is records missing residential address information and therefore a Meshblock was unable to be assigned during geocoding. Extensive analysis following any linkage would be required to ascertain the characteristics of temporary migrants in both the linked and unlinked records in order to effectively account for any systematic bias in the linked data.

The assessment of the feasibility of undertaking future linking of the TVH records to the 2011 Census concluded that, with simulated linkage results of a 70% link rate and a 98% precision rate, all the required elements are in place to produce a useful dataset for analysis. This is particularly true for the largest groups of temporary migrants (International Students and Temporary work (skilled) subclass 457). However, in order to improve the link rate, and provide more reliable and more detailed estimates, more work needs to be done to assess and improve data quality and address any issues that arise from the under-representation of certain subgroups in the linked dataset.

The study also found that data linkage of the TVH records to the ATO PIT data is not feasible, primarily due to the quality of the address data. There are very few variables common to both datasets, so successful linkage to PIT data requires high quality name and address data and the Temporary migrants dataset was not of sufficient quality due to the lack of high quality address information for all records.

ABBREVIATIONS

ABS	Australian Bureau of Statistics
ACMID	Australian Census and Migrants Integrated Dataset
ACTMID	Australian Census and Temporary Migrants Integrated Dataset
ANZSCO	Australian and New Zealand Standard Classification of Occupations
ASB	Analytical Services Branch, ABS
ASCO	Australian Standard Classification of Occupations
ATO	Australian Taxation Office
CDE	Census Data Enhancement
DI	Data Integration
DIBP	Department of Immigration and Border Protection
DHS	Department of Human Services
DIMIA	Department of Immigration and Multicultural and Indigenous Affairs
DSS	Department of Social Services
ICSE	Integrated Client Services Environment
NMSU	National Migrants Statistics Unit
PIT	Personal Income Tax
SACC	Standard Australian Classification of Countries
SDB	Settlement Database
TRIPS	Travel and Immigration Processing System
TVH	Temporary Visa Holders

ACKNOWLEDGEMENTS

This paper was prepared by the ABS National Migrant Statistics Unit (NMSU), with assistance from the ABS Analytical Services Branch (ASB). 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 study "Assessing the Suitability of Temporary Migrants Administrative Data for Data Integration" was funded by DIBP and the results are based on temporary migrant administrative data supplied by the DIBP to the ABS. Any discussion of data limitations or weaknesses is in the context of using the data for statistical data integration purposes, and is not related to the ability of the data to support the DIBP'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 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 assistance of the ABS Analytical Services Branch, in particular Peter Rossiter for his work in conducting the linkage simulations and compiling the results, which are reported in Section 4.1 and the Appendix.

CONTENTS

APPENDIX

ASSESSING THE SUITABILITY OF TEMPORARY MIGRANTS ADMINISTRATIVE DATA FOR DATA INTEGRATION

National Migrants Statistics Unit
Australian Bureau of Statistics

ABSTRACT

In 2014, the Australian Bureau of Statistics was provided with access to the Department of Immigration and Border Protection's Temporary Visa Holders (TVH) records to assess their suitability for data integration, in particular with the 2011 Census of Population and Housing. This suitability was ascertained by analysing the quality of the TVH data items as well as an assessment of the linking feasibility with a simulated linkage study. The study concluded that the TVH data were suitable for linkage with the Census, however extensive analysis would need to be conducted on the linked and unlinked files to ensure an adequate representation of the TVH and the development of an effective calibration strategy. The study also found that linking the TVH records to the Australian Taxation Office (ATO) Personal Income Tax (PIT) data is not feasible, primarily due to the quality of the address data.

1. INTRODUCTION

The aim of this study is to assess the quality of the Department of Immigration and Border Protection's (DIBP) Temporary Visa Holders (TVH) file in terms of its suitability for data integration. It investigates methodologies to improve future data integration studies and compares the quality with earlier quality studies linking the 2011 Census and the Australian Government Settlement Database (SDB) (which only includes permanent migrants). The feasibility of linking the TVH file with records from the Australian Bureau of Statistics' 2011 Census of Population and Housing without the use of a unique record identifier to create a new dataset for statistical and research purposes is also assessed by utilising a linkage simulation tool being trialled by the ABS. The feasibility of linking the TVH data to the Australian Taxation Office (ATO) Personal Income Tax (PIT) unit record data is also discussed.

The Australian Bureau of Statistics (ABS) has strong safeguards in place to protect identifiable information such as name and address. 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 a way that enables identifiable information to be associated with a specific person.

Assessing the suitability of Temporary Migrants Administrative Data for Data Integration has three steps:

Step 1: Assessment of the data quality

This involves looking at the quality of the data items on the TVH dataset in terms of completeness, comparability and an analysis of the characteristics of the data and the usefulness of a linked dataset for evidence-based policy and research. Results are provided in Section 3.

Step 2: Assessment of the linking feasibility

This involves an assessment of the feasibility and quality of linking the TVH data to the 2011 Census data file by simulating linkage using an experimental linkage tool. Results are summarised in Section 4.1, with a detailed technical report provided in the Appendix.

Step 3: Assessment of overall feasibility

Taking into account the results of the assessment of data quality and linking feasibility (Steps 1 and 2), judgements can then be made about whether a resulting linked data file will be suitable to support relevant analysis of the characteristics of various visa classes of temporary migrants. This conclusion is provided in Section 5.

2. THE POTENTIAL OF A NEW AUSTRALIAN CENSUS AND TEMPORARY MIGRANTS INTEGRATED DATASET (ACTMID)

The research and policy issues identified by the NMSU through ongoing consultation with various stakeholders on their needs and priorities as they relate to temporary migrants and which could potentially be addressed with ACTMID include:

- economic contributions of temporary migrants (e.g. through income and labour force participation);
- geographical distribution of temporary migrants (at least to Capital City/Rest of State and Statistical Area 4 level);
- housing types – for example, privately owned, rental and shared accommodation;
- the extent to which temporary work (skilled) (subclass 457) visa holders are in jobs commensurate with their skills (i.e. educational qualifications);
- employment experience, qualifications, employment barriers (e.g. proficiency in English) and whether labour market skills gaps are being met;
- the success of temporary overseas students in the labour market compared with recent permanent resident graduates and Australian born graduates;
- exploratory cross-sectional analysis of temporary and permanent migrants in conjunction with the Australian Census and Migrants Integrated Dataset over time.

Information on the economic contributions of temporary visa holders will be available through variables that collect information such as employment status, occupation and total personal weekly income from the Population Census. Geographical distribution will be available through a variety of different data items that collect place of enumeration, usual residence and place of work at a number of different sub-state geographic levels.

Person level data is also available on the Census for usual address one year ago (and five years age) which would provide some information on the patterns of movement of temporary migrants after arrival. In addition, SEIFA data are collected.

It would be possible to analyse whether the jobs held by temporary visa holders are commensurate with their educational qualification levels. The Census collects information on the highest qualifications held and highest year of school completed, as well as occupation in their current job. It would be possible to compare the Census education and occupation outcomes for all migrants and the Australian born population with the TVH data in order to analyse differences in occupation and educational attainment.

Subject to confidentiality, information on housing type (e.g. privately owned, mortgaged or rental accommodation) is also available on the Census. However, for the 2011 ACMID linking project, this information presented a higher risk of identifying particular individuals and families and so was not included in the final output file. Consequently, a TVH file linked to the 2011 Census may pose a similar risk, and may be subject to the same conditions.

Longitudinal analysis of migrants is an important and recognised data need. Future longitudinal analysis of temporary migrants, including their movement from temporary to permanent residency is possible. However, such analysis would be dependent upon the establishment of a longitudinal ACTMID in conjunction with a longitudinal ACMID (i.e. Australian Census linked to both temporary and permanent migrants). If a TVH record was no longer present on a subsequent file, it would be easy to ascertain whether they had been granted a permanent visa as they would be present on the SDB file of permanent migrants with the same Travel and Immigration Processing System (TRIPS) personal identifier.

3. ASSESSMENT OF THE QUALITY OF THE TEMPORARY MIGRANTS DATA

Section 3.1 provides an overview of the Temporary Visa Holders (TVH) file, and Section 3.2 discusses the quality of the data items on the TVH file. Section 3.3 identifies potential issues of representativeness that may arise from the linkage process. Section 3.4 examines the potential for key linking variables to uniquely identify individuals in the TVH file.

3.1 Temporary Visa Holders (TVH) file

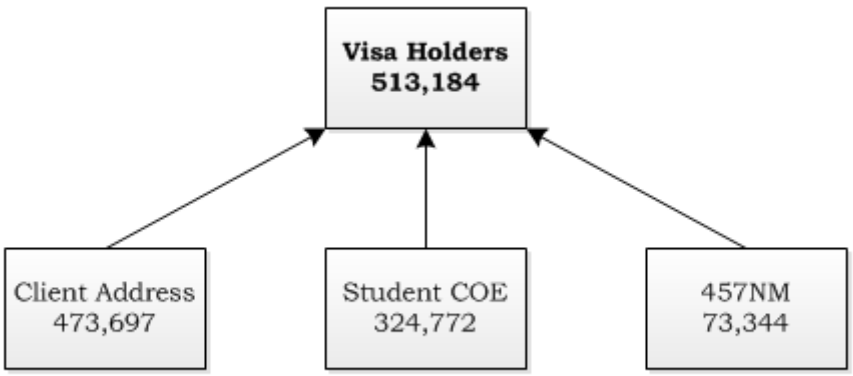
The TVH file is compiled by the Department of Immigration and Border Protection (DIBP) from various departmental administrative systems and a number of external sources. It contains the records of all persons in Australia holding a temporary visa. These visas include work visas, study visas, holiday visas and many others. The dataset utilised for this study is based upon International Student and Temporary work (skilled) (subclass 457) visa holders in Australia as at 31 July 2011. This is because the majority of other temporary visas classes are quite small. There are large numbers of people on Working Holiday visas but a minority of these would be counted as usual residents in Australia on Census night. (On the Census form, those persons who usually live in another country and who are visiting Australia for less than one year mark "Other country" and are identified as "Overseas visitors"). The date of 31 July 2011 was selected as it is the closest available to the Census date of 9 August 2011.

The Temporary Visa Holders file is comprised of the following four extracts:

- Visa holders,
- Client Address,
- Student Confirmation of Enrolment (COE), and
- Temporary work (skilled) visa nominations (457NM).

The largest extract is the Visa Holders file with 513,185 temporary visa holder records. However one record was identified as a duplicate and removed, leaving 513,184 records. The additional data on the other three extracts could be linked to the Visa Holders records using identifiers. All of the potential linking variables (see Section 3.2) were contained on either the Visa holders file or the Client Address file. The Student COE and the Temporary work (skilled) visa nominations (457NM) files contained primarily analysis variables e.g. sponsor industry and nominee occupation.

3.1 Temporary Visa Holders file



3.1.1 Extract 1: Visa Holders

This file contains demographic information on the temporary visa holder such as name, date of birth, sex, marital status, country of birth and visa subclass based upon the Travel and Immigration Processing System (TRIPS) visa manager.

The following visa subclasses are present on the file:

- Temporary work (skilled) (subclass 457);
- International student visas:
    - Independent English Language Course for Overseas Students (ELICOS) (subclass 570);
    - Schools Sector (subclass 571);
    - Vocational Education and Training Sector (subclass 572);
    - Higher Education Sector (subclass 573);
    - Postgraduate Research Sector (subclass 574);
    - Non Award Sector (subclass 575); and
    - AusAID or Defence Sponsored Sector (subclass 576).

This information was obtained from two source systems (DIMIA, 2005):

- Offshore visa applications are processed using the Immigration Records Information System (IRIS); or
- Onshore visa applications are processed using the Integrated Client Services Environment (ICSE).

3.1.2 Extract 2: Client Addresses

This file contains information on the address of the visa holder. This information is recorded as a residential address, a business address or a postal address. Issues with address are discussed in more detail in Section 3.2 of this paper.

The addresses may be located in Australia or overseas.

3.2 Type of addresses by location

	Residential		*Business*		*Postal*		*Total (a)*	
	No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
Location of address								
In Australia	362,013	81.03	3,405	16.02	254,636	76.25	362,013	76.42
Overseas	84,748	18.97	17,848	83.98	79,302	23.75	111,684	23.58
Total	446,761	100.00	21,253	100.00	333,938	100.00	473,697	100.00

(a) Total may not add to the sum of components as some records may have more than one type of address.

Visa and sponsor address details are for clients recorded in ICSE only. There is no opportunity to record addresses for records in the IRIS source system. However, there are some records where the visa application was processed in IRIS but was recorded in ICSE. Therefore, these records may have address information. Consequently, 11% of the Visa holders file did not have a corresponding client address record.

Address records were geocoded to the Meshblock and Statistical Area 1 levels of the *Australian Statistical Geography Standard* (ASGS)(ABS, 2014) by the Geography Section of the ABS in order to create a linking variable appropriate for Bronze Standard linkage.[1]

3.1.3 Extract 3: Student Confirmation of Enrolment (COE)

The COE file contains information on the level of education currently being undertaken by the individual as well as the name of the institution at which the individual is enrolled.

The student data is based upon the primary or main applicant student visa holder's COE details. Where more than one enrolment confirmation is recorded on the visa application, the enrolment confirmation with the highest educational level is recorded. AusAID or Defence Sponsored Sector visa holders (subclass 576) do not have an enrolment confirmation.

1 Bronze standard linkage implies that name and detailed address information are not used as linking variables.

3.1.4 Extract 4: Temporary work (skilled) visa nominations (457NM)

This file contains information for migrants holding a Temporary work (skilled) visa (subclass 457). It also contains some information on their sponsors. Holders of this visa may stay in Australia and work for an approved business for up to four years duration. Records for subclass 457 nomination (NM) visa holders may contain information on the occupation of the visa holder, the industry of the employer and the total remuneration or guaranteed earnings received by the subclass 457 Primary applicant.

The 457NM data is based on the latest data from the subclass 457 visa application of the Primary or main applicant. Where more than one nomination is recorded for the subclass 457 visa application, the latest entry is provided.

The total remuneration data in the nomination application is the gross earnings the employer expected to pay the subclass 457 primary visa holder, which includes any fringe benefits. This figure is based on the time the nomination was approved, which may be up to four years prior to the date of extraction (31 July 2011). In practice however, it should be noted that the subclass 457 visa holder may receive a higher rate of remuneration due to salary indexing each year due to the Consumer Price Index and/or keeping up with the Temporary Migration Income Threshold.

The nominated occupation is classified according to the *Australian and New Zealand Standard Classification of Occupations* (ANZSCO) (ABS, 2013a). ANZSCO coding of occupation data was introduced in the DIBP on 1 July 2010. Applications lodged prior to that date are classified according to the *Australian Standard Classification of Occupations* (ASCO) (ABS, 1997). These records were coded to an ANZSCO code using a standard DIBP concordance approved by the ABS.

3.2 Assessment of the variables for linking

3.2.1 Name

Name data on the TVH file was of good quality. Of the 513,184 records on the file, none had a missing surname and only 1.4% (6,935) had a missing given name. Experience gained from the quality study for the ACMID linking project found that countries such as India and Indonesia are more likely to report both the individual's given names and surname in the one field. Examination of the TVH data showed that of the 6,935 records with a missing given name, 72% (5,018) had more than one name in the surname field. These records were most likely reporting a given name in the surname field. The concatenation of first name and surname in the surname field can be addressed prior to linking as a routine name repair. Names were considered to be of sufficient quality to be used for gold standard linking.

3.2.2 Addresses

The following issues have been identified with the address information provided for temporary migrants for this study. Whilst these issues have implications for the number and quality of the geocoded addresses for linking, and hence the accuracy of the linkages, they do not necessitate the removal of any of the records.

- Address details are based on clients that are recorded in ICSE only. There is no opportunity to record addresses in the IRIS source system. However, there are some records where the visa application was processed in IRIS but the applicant had previously had another case recorded in ICSE.
- The address data may be an onshore or offshore address. An offshore address cannot be linked to a Census record as the Census record has the place of usual residence in Australia.
- The address data may not be that of the temporary visa holder (e.g. a Migration Agent's address).
- The address data may be missing, incomplete or not up to date.

For the ACMID linking project (Richter *et al.*, 2013), address information on the Settlement Database (SDB) was regularly updated by Medicare.

Australia has signed Reciprocal Health Care Arrangements (RHCA) (DHS, 2014) with the following countries:

The United Kingdom,
The Republic of Ireland,
New Zealand,
Sweden,
The Netherlands,
Finland,
Belgium,
Norway,
Slovenia,
Malta, and
Italy.

These agreements entitle new arrivals and visitors to Australia from these countries to restricted access to emergency medical cover in Australia. These agreements allow visitors to enrol for Medicare under a Reciprocal Health Care Agreement. However, students from Norway, Finland, Malta and Ireland are not covered by an RHCA.

Only 8.9% of TVH records are from eligible countries and as there is no onus on these migrants to register for Medicare, the proportion of those who do register may be considerably less. In terms of the cost and effort involved, it is not considered viable for DIBP to pursue updates for such a small cohort. Since over 90% of the migrants on the TVH file were not eligible for Medicare, their address information cannot be updated in this way. For this reason the TVH addresses have a higher potential than the addresses of permanent visa holders on the SDB to be out of date or inaccurate in some way.

Repeating addresses

There were several instances of the same address being reported multiple times on the Client Address file. The most commonly reported Residential addresses appeared to be student accommodation e.g. university or college villages. Some were offshore.

For example, over 100 temporary migrants had an identified migration and recruitment agency in South Australia as their address. There were four university villages with more than 50 temporary migrants reporting those addresses. The most commonly reported address was for an offshore (Malaysia) migration agent specialising in international education.

It is worth noting that an address may be reported many different ways and actual counts of commonly reported addresses may be higher than the figures quoted above.

Some TVH addresses were not for the individual concerned but were for a Migration Agent instead. These Migration Agents tended to report a postal address and not a residential address. As postal addresses, in particular Post Office Boxes, cannot be geocoded to a Meshblock with the same accuracy as an actual street address they are not as useful for linking. In total, 4.4% of all records on the Client Address file reported a Post Office Box in at least one of the address categories.

Geocoding of addresses

Address information needed to be geocoded to obtain the Meshblock and Statistical Area 1 (SA1) levels of geography (ABS, 2014). Due to the requirement for compatibility with the Census usual address to match records from the TVH, the only addresses that were submitted from the Client Address file for geocoding were those with an Australian residential address. Offshore addresses, postal and business addresses were excluded as they were unlikely to be useful for establishing a link. The usefulness or otherwise of the postal and business addresses for linking can be further investigated and ascertained if linking and a quality study are undertaken.

It should be noted that some records with a response for 'Residential address' contained addresses that did not appear to be a residential address. There were some instances where business addresses and P.O. Boxes were reported as a 'Residential address'.

Approximately 364,000 records on the Client Address file had an Australian residential address and were submitted for geocoding. After geocoding, 94% of the addresses had been successfully coded to a Meshblock and 95% to SA1 level.

This equated to 67% of the records on the TVH file having a valid code for Meshblock and SA1. As address information was only considered of sufficient quality for the residential addresses, this meant that a large proportion of the records (33%) were not geocoded to Meshblocks. This will have a significant impact on the number and quality of links generated.

Analysis of those records that did not code to Meshblock showed that approximately 60,000 of them, or around 35% of the non-geocoded records, were Secondary applicants.

For International Students it is possible to identify a family unit (and hence Secondary applicants) by the case ID where the applications are made at the same time. Similarly, Secondary applicants on subclass 457 visa applications are attached to the same case ID as the Primary applicant. Consequently, where an address is missing for a Secondary applicant but is present for the Primary applicant with the same case ID, the same address could be assigned to the Secondary applicant record prior to geocoding. This would mitigate the issue of missing address and improve the number of records successfully assigned a Meshblock for linking. There were 20,000 Secondary applicants identified without an address that had the same case ID as a Primary applicant with a residential address. If the address details of the Primary applicant were utilised for these Secondary applicants, a further 3% of addresses would be geocoded and 70% of the TVH file would have a Meshblock assigned. Note that this was not done for the TVH file used in the simulated linking analysis.

There were 17 Visa Application numbers that had more than one Primary applicant. These were not duplicate records and some appeared to be couples both listed as the Primary applicant living at the same address.

Quality of the Meshblock coding for Temporary Visa Holders addresses

When the TVH address information was geocoded to Meshblock, the results included a diagnostic report which is based upon whether any address issue was encountered by the Address Coder in the coding process. The Address Coder is able to make some amendments to records during the coding process based upon the information provided in the address record.

Where an address is amended or repaired in order to match an address in the coder, the amendment is given a 'risk' rating that indicates a level of risk associated with the accuracy of the results that are returned. For example, if the coder added a missing postcode or amended a spelling mistake in the suburb name the record would be considered 'low' risk. However if a missing state or territory code were added in response to partial address information (e.g. a street name or suburb) the record would receive a 'high' risk rating because that street name or suburb may exist in more than one state or territory and the correct one may or may not have been selected.

The results of the risk analysis are detailed below.

3.3 Geocoding risk analysis

Risk level	*Recommended action*	*Number*	*Percentage*
Not amended	Accept	207,044	57.2
Low	Accept	150,576	41.6
Medium	Review	3,810	1.1
High	Reject	573	0.2
Total		362,223	100.0

Learnings from the 2011 ACMID linking project showed that it was acceptable to include Meshblocks from the Low and Medium risk levels for linking as the records were generally of sufficient quality. For the TVH, if only those records that had a high risk level Meshblock were rejected this would only result in a loss of 0.2% of the records.

Options for dealing with the missing address information

In order to deal with the issue of missing address information, a number of strategies could be undertaken.

(a) Option 1 – Conduct linkage with current TVH file

The linkage rate of the 2011 ACMID linkage project was considered feasible at 76%. (See Richter *et al.*, 2013 for details.) If results from the linkage simulation tool indicate that a similar linkage rate is achievable with the TVH file (irrespective of any issues with the lack of residential addresses for some cohorts of temporary migrants), it is likely that linking the current TVH file to the 2011 Census file is feasible.

(b) Option 2 – Link only records with a value for Meshblock

Geocoding to Meshblock is only achievable with good quality residential addresses (with a street name and usually a number). For this reason it may be more appropriate to conduct the linkage process using only those records with a valid code for Meshblock.

Any under-representation or over-representation within the resultant integrated dataset would be identified in the subsequent quality study of the integrated dataset. The calibration or 'weighting' of the integrated dataset would compensate for the component of the temporary migrants population without a valid Meshblock and seek to address any under-representation or over-representation present in the integrated dataset by benchmarking to known temporary visa holder subpopulations. The tables below provide an indication of the under and over-representation that may result from the presence or otherwise of a Meshblock code for records.

The following subpopulations will be under-represented in the Meshblock only file:

- Migrants aged less than 15 years;
- Migrants born in Bhutan, Denmark, Oman and Norway (and, to a lesser extent, North West Europe);
- Secondary applicants; and
- Temporary work (skilled) (subclass 457) visas.

The following subpopulations will be over-represented in the Meshblock only file:

- Migrants aged 25–34 years;
- Migrants born in Nepal and Lebanon (and, to a lesser extent, Southern and Central Asia);
- Primary applicants; and
- Vocational Education and Training Sector (subclass 572) visas.

Table 3.4 shows that there is little difference in the Original file and Meshblock only file in the proportion of TVH records by age group, with the exception of those aged less than 15 years and those aged 25–34 years.

3.4 Relative frequencies (%) in each Age group, for the original TVH file and the Meshblock only TVH file

Age group	*Original TVH file*	*Meshblock only file*
Under 15 years	8.3	4.1
15–24 years	42.0	42.4
25–34 years	36.6	41.0
35–44 years	9.7	9.6
45–54 years	2.6	2.3
55–64 years	0.6	0.5
65 years and over	0.1	0.1

In table 3.5 there is little difference in the proportion of subpopulations by region of birth on the file with Meshblocks.

3.5 Relative frequencies (%) in each Region of birth, for the original TVH file and the Meshblock only TVH file

Region of birth	*Original TVH file*	*Meshblock only file*
Oceania and Antarctica	1.8	1.5
North West Europe	12.1	10.4
Southern and Eastern Europe	1.9	1.8
North Africa and Middle East	4.6	4.2
South East Asia	18.8	17.7
North East Asia	28.5	29.8
Southern and Central Asia	20.0	23.3
Americas	7.7	6.6
Sub-Saharan Africa	3.8	3.8

Analysis of the proportions of visa holders by country of birth, rather than region of birth, gives a better indication of the over-representation and under-representation. Table 3.6 displays the greatest changes in proportion for countries from which Australia received at least 500 migrants. For example, migrants born in Bhutan will be under-represented on the Meshblock only file by 65% while migrants born in Nepal will be over-represented by almost 31%.

3.6 Number and relative frequencies (%) in selected countries of birth, for the original TVH file and the Meshblock only TVH file and percentage change of relative frequency

	Original TVH file (513,184 records)		*Meshblock only file (341,761 records)*		*% change of relative frequency*
Country of birth	*No.*	*%*	*No.*	*%*	
Most under-represented					
Bhutan	569	0.11	133	0.04	–64.9
Denmark	901	0.18	247	0.07	–58.8
Oman	607	0.12	178	0.05	–56.0
Norway	1,541	0.30	491	0.14	–52.2
United States of America	13,284	2.59	4,988	1.46	–43.6
Austria	553	0.11	235	0.07	–36.2
Papua New Guinea	2,018	0.39	876	0.26	–34.8
Netherlands	1,986	0.39	877	0.26	–33.7
Belgium	529	0.10	240	0.07	–31.9
Kuwait	861	0.17	393	0.11	–31.5
Most over-represented					
Nepal	14,159	2.76	12,315	3.60	30.6
Lebanon	943	0.18	813	0.24	29.5
Egypt	958	0.19	780	0.23	22.3
Colombia	5,189	1.01	4,220	1.23	22.1
Peru	1,370	0.27	1,106	0.32	21.2
Pakistan	7,534	1.47	6,065	1.77	20.9
Mauritius	3,123	0.61	2,437	1.77	17.2
Kenya	1,742	0.34	1,350	0.40	16.4
Ireland	8,372	1.63	6,453	1.89	15.7
Iran	3,234	0.63	2,490	0.73	15.6

In table 3.7, primary applicants with a Meshblock had a higher representation at 85% while secondary applicants with a Meshblock had an almost 7% lower representation after coding to Meshblock.

3.7 Relative frequencies (%) in each Applicant type, for the original TVH file and the Meshblock only TVH file

Applicant type	*Original TVH file*	*Meshblock only file*
Primary	78.3	85.2
Secondary	21.7	14.8

In table 3.8, the largest differences in the proportion of subpopulations on the file with Meshblocks were for Temporary subclass 572 and 457 visa holders.

3.8 Relative frequencies (%) in each Visa subclass, for the original TVH file and the Meshblock only TVH file

Visa subclass	*Original TVH file*	*Meshblock only file*
457	26.3	22.8
570	2.8	2.5
571	3.8	3.5
572	18.5	23.0
573	42.0	44.2
574	3.5	2.8
575	1.8	0.7
576	1.2	0.4

(c) Option 3 – Subset the TVH file further to only those records with a high degree of accuracy for their Meshblock code

In order to get just the highest quality addresses, it is possible to subset the TVH file to use just the records where the geocoded value had a risk level of Not amended, Low or Medium.

As shown in table 3.3, this would only exclude 0.2% of the file.

3.2.3 Date of birth and derived age

Date of birth is recorded on the TVH file as day, month and year of birth. Age as at Census night was easily calculated for this study.

There were only 36 records with a missing value for day of birth and of those, 35 also had a missing month of birth. None of the temporary visa holders had a missing value for year of birth, with these values ranging from 1907 to 2012. It is most likely that individuals holding a student or a Temporary work (skilled) subclass 457 visa in the older or very young age groups are Secondary applicants.

3.9 Relative frequencies (%) of Temporary visa holders, by age group, applicant status and visa type

	Primary applicant		*Secondary applicant*	
Age group	*Temporary work (skilled)*	*International student*	*Temporary work (skilled)*	*International student*
Under 15 years	0.1	0.6	41.4	30.4
15–24 years	3.7	61.3	10.0	10.9
25–34 years	55.1	33.3	24.4	46.0
35–44 years	28.1	4.1	16.7	10.8
45–54 years	9.9	0.5	5.6	1.7
55–64 years	2.8	0.1	1.4	0.2
65 years and over	0.3	0.0	0.3	0.1
Total	100.0	100.0	100.0	100.0

During the ACMID data linking project it was evident that where a visa applicant only knows the year in which they were born and is unable to provide a month or day of birth, they are sometimes allocated an 'administrative date' for their date of birth. These dates are:

- January 1,
- July 1, and
- December 31.

"January 1" was the most frequently reported day of birth in the TVH file, however this does not appear to have affected more than approximately 1,000 records. Additionally, "July 1" and "December 31" appear no more frequently than other dates, indicating that these 'administrative dates' are not a large issue on the TVH file.

3.2.4 Marital Status

The following categories were available on the TVH file for Marital Status:

- De facto partner
- Divorced
- Engaged
- Married
- Never married
- Not stated
- Separated
- Widowed

Marital status was recoded to align the categories with the Census data item for Registered Marital Status. The categories are as follows:

- Not applicable (aged less than 15 years)
- Never married
- Married
- Divorced
- Separated
- Widowed

Visa holders on the TVH file under the age of 15 on 31st July 2011 were recoded to "Not applicable" to be comparable with Census. It is more problematic to recode the categories of "Engaged" and "*de facto* partner". This is because there is no way to determine whether these people have ever been married, separated, divorced or widowed prior to their engagement or entering into a *de facto* relationship. However, for the purposes of this study these persons are coded to "Never married".

3.2.5 Country of birth

The country of birth and citizenship country of individuals recorded on the TVH file were concorded from four-character codes to four-digit numeric codes from the *Standard Australian Classification of Countries* (SACC) (ABS, 2011).

There were 4,580 temporary visa holders that listed their country of birth as Australia. The majority of these migrants were Secondary applicants on the visa application (97%). However, 3% were listed as the Primary applicant. No records had recorded "Australian" as their citizenship, which is to be expected. All were citizens of another country.

In some cases, the information provided for Country of birth can be insufficient for four-digit SACC coding (especially for migrants from the United Kingdom), or may identify a country that no longer exists (e.g. the former Soviet Union or Former Yugoslavia). However, this information may be sufficient to assign a two-digit or one-digit SACC code.

Country of citizenship may or may not reflect the country in which the person was born, but this information could be used to impute a four-digit country of birth code for those applicants whose responses are missing. However, more analysis would be required after linking to determine if this was worthwhile or caused problems.

On the TVH file, 1885 records had a value for country of birth as "Unknown". After concording the countries of birth to SACC, 1890 records were coded to 'Inadequately described' due to insufficient information to adequately code them to a country of birth.

3.2.6 Year of arrival

The Census Year of arrival variable was compared with the closest equivalent on the TVH file (i.e. Date of grant) for the purposes of this study. However, it would be possible to obtain the date of first arrival and the most recent arrival from TRIPS for comparison in any future linkage work. The Census year of arrival is the year the person first arrived in Australia to live for one year or more. Whilst this first arrival and most recent arrival data may also prove to not match up exactly, some latitude in the degree of accuracy would be acceptable, such as ± one year. Whilst these Arrival dates were not supplied for this study, they are likely to provide a better approximation than the variable that was used as a proxy for year of arrival in this study, i.e. the date in which the temporary visa application was granted.

3.2.7 Other variables required

In order to remove out of scope records from the file prior to linking, several other variables would be required on the TVH file. Out-of-scope people would include those who had passed away prior to Census night or who were not in Australia on Census night.

Deaths

Some benefit may additionally be gained if deceased persons were flagged on the Client file to enable the identification of records for people who have died prior to Census night. If this information proves to be sporadic in nature then the number of deaths would need to be accounted for utilising ABS Demography data in the calibration process after linking.

Absences on Census night

TRIPS information that would allow the identification of Temporary visa holders who were not in the country on Census night would be a welcome addition to the TVH file as it would enable those persons who were overseas and had not filled out a Census form to be excluded prior to linking.

3.3 The requirement for a calibration strategy for a linked file

No linkage process involving the TVH file will achieve a 100% linkage rate. As a result, a calibration process will need to be undertaken for the linked dataset to be more representative of the International student and Temporary work (skilled) subclass 457 population.

If the linkage process is run on only a proportion of the TVH file, for example only those records with a value for Meshblock, extra care will need to be taken to ensure that the calibration process accounts for any potential bias (over and under representation of subgroups) created by the removal of records from the file as well as the unlinked records after linking to the final analysis file. A quality study will be required to ascertain the characteristics of the linked and unlinked temporary migrants and the calibration strategy will need to account for those missed, under-represented or over-represented.

This strategy will have more complexity than the 2011 ACMID linking project, where permanent migrants were generally expected to have been included in the 2011 Census. Temporary migrants are less likely to report on the Census as they may consider it applies only to permanent residents. Young males from 19 to 25 years of age are generally harder to capture in the Census irrespective of residency status or country of birth and this could potentially affect outcomes for the International students' cohort.

It should be noted that missing values in the address fields do not appear to be random. There may be differences in the characteristics of the offshore records processed through the IRIS source system compared with the onshore records processed through the ICSE source system. It is also possible that whole groups of temporary migrants with the same characteristics (e.g. country of birth) would be missing if they have an offshore address on their migration application or that of a Migration Agent onshore. Issues such as these require further investigation.

The unlinked or missing component of the temporary migrant population would need to be successfully modelled in order to calibrate the linked component to the total temporary migrant population. This would pose significant challenges in terms of representation of the diverse groups from many different countries of birth that would be present in the data for which there may be little or no additional information. At present there does not appear to be a way for DIBP to improve the quality of address data as this is done via Medicare updates for permanent migrants and temporary migrants are not eligible for Medicare.

3.4 Some analysis of combinations of linking variables

International students may be more likely to be located close to large educational institutions in certain areas (i.e. student housing or suburbs close to universities with affordable rents) and tend to be within a particular age range (15–34 years). For this reason there is a risk of a much higher rate of multiple potential matches that could arise for each individual. Therefore, the records on the TVH were analysed for the frequency of certain combinations of characteristics.

To ascertain the frequency of records with the same information on key linking variables, the records on the Visa holders file were given a linkage identifier for the purposes of this study which was created by concatenating the information on the following candidate linking variables:

- Country of birth (SACC)
- Date of birth
- Marital status (concorded to Census)
- Sex
- Meshblock

These identifiers were then examined for any repetition. Of the 513,184 records on the Visa holders file, there were 480,727 unique linkage identifiers. These identifiers were present on up to 14 records with the following frequencies:

3.10 Frequency of linking identifiers on the Temporary Visa Holders file

# of records with a given identifier	*Frequency*	*Cumulative frequency*	*Percent*
1	461,517	461,517	96.00
2	12,626	474,143	2.63
3	3,414	477,557	0.71
4	1,469	479,026	0.31
5	777	479,803	0.16
6	464	480,267	0.10
7	234	480,501	0.05
8	111	480,612	0.02
9	67	480,679	0.01
10	29	480,708	0.01
11	11	480,719	0.00
12	4	480,723	0.00
13	2	480,725	0.00
14	1	480,726	0.00

For example, 461,517 identifiers were present on only one record, indicating that they were unique on the dataset. However, 12,626 identifiers were present on two records, meaning that this combination of information on the linking variables was not unique and so two records would potentially link to all the same Census records, with no way to distinguish between them.

In the most extreme example given here, one identifier was present on 14 records, so these 14 records would potentially link to all the same Census records. Even if these records generated only 14 potential Census records, there would be no way to determine which TVH record matched which Census record.

It should be noted that the majority of these repeating identifiers had a missing value for Meshblock which reinforces the importance of this linking variable for a good linking outcome. Where a valid code for Meshblock was present, 99.9% of identifiers were present on only one record.

4. ASSESSMENT OF THE LINKING FEASIBILITY

4.1 Simulated linking

Statistical data integration utilising probabilistic record linkage has been successfully undertaken by the ABS on a number of occasions (e.g. ACMID, ACLD). The ABS has developed an experimental software application to facilitate the statistical simulation of outcomes from probabilistic record linkage. The application is designed to inform feasibility analyses of proposed record linkage projects, and quality analyses of completed linkage projects. The theoretical basis for the statistical simulation approach is described in Chipperfield and Chambers (2015).

By utilising a combination of factual information and hypothetical (but informed) assumptions, it is possible to estimate the proportion of records likely to be linked, and the link accuracy (or precision) of those links. The decision regarding whether to proceed with the actual record linkage can be then informed by the prediction that link accuracy will be:

- high (greater than 95% links correct),
- moderate (75–95% links correct), or
- low (less than 75% links correct)

at an acceptable link rate.

The simulation study conducted for this project concluded that:

> *There is potential to link 70% of the records on the Temporary Visa Holders file to the 2011 Census with a precision (or link accuracy) of 98% or higher.*

A higher link rate may be achieved by trading off link accuracy.

Linked records that agree on Meshblock are more likely to be accurate, and it would be unsafe to assume that the average link rate and link accuracy can be applied to all subpopulations of the TVH dataset.

For further technical details of the assessment using the linkage simulation tool, please see the Appendix, *An Assessment of the Feasibility of Linking Temporary Visa Holder Records to the 2011 Census*.

4.2 Feasibility of linking Temporary Visa Holders data to the 2011 Census and Personal Income Tax data

4.2.1 2011 Census of Population and Housing

There are a variety of data items on the TVH file that could be used as candidate linking variables for linkage to the 2011 Census. However, it is clear from all of the analysis conducted above that the presence of high quality onshore residential addresses that can be successfully coded to Meshblock provides significant improvements in the linkage process. The fact that only 67% of records on the TVH file could be coded to Meshblock will lead to a sub-optimal linking outcome. Note that it is possible to improve this figure to 70% if Primary applicant addresses are utilised for Secondary applicants with the same case ID (see section on Geocoding of addresses).

The results from the linkage simulation study discussed in Section 4.1 (and the Appendix) may represent the higher end of the spectrum in terms of linking outcomes. The linking results could be considerably lower if certain subpopulations of Temporary migrants fail to respond to the Census, irrespective of their eligibility.

4.2.2 Australian Taxation Office's Personal Income Tax (PIT) file

As discussed in the research paper *Feasibility Study of Linking Migrant Settlement Records to Personal Income Tax Data* (Walsh and Weckert, 2014), any linking project undertaken utilising the Personal Income Tax (PIT) administrative data would need to be Gold Standard probabilistic linking, with names and addresses as linking variables. This is primarily due to the presence of few other variables for linking on the PIT file.

As discussed in Section 3.2, Given name and Surname are considered to be of sufficient quality for name information to be utilised for linking. However, as also mentioned in Section 3.2, address information is not of sufficient quality to enable comparison with PIT record addresses for a large proportion of records. These issues with addresses mean that Gold Standard linking could not be utilised for these records. However, should employer data become available from the ATO in the future, this may provide some additional linking variables with similar or greater potential than address information for linking.

Without the benefit of address information, it is unlikely that good linkage results would be obtained. Variables such as Marital Status, Country of birth and Year of arrival are not available on the PIT file and cannot be used as linking variables. Whilst it may be possible to conduct the linkage process using just those records with adequate address information, it would be difficult to effectively calibrate the linked data file to account for those records not included because the number of temporary visa holders, especially students, who have not submitted a tax return in a given year, would be very difficult to estimate.

5. CONCLUDING REMARKS ON OVERALL FEASIBILITY

An assessment of the feasibility of undertaking future linking of the TVH records to the 2011 Census and PIT records requires a judgment about whether the quality of any resulting future linked data set (ACTMID) would be fit for purpose. That is; given a link rate of around 70% (or about 350,000 temporary migrant records) and the need to resolve issues with data quality, would a resulting ACTMID file be useful for analysis of the temporary migrant population?

The most important benefit of linking the TVH file to the Census is the ability to undertake analysis by the specific *visa subclass* of the temporary migrant. In this study linking has only been attempted for the two largest visa subclasses – International Students and Temporary work (skilled) (subclass 457) visas. Assuming a successfully calibrated file, output from a linked 2016 Census data set may include approximately 350,000 International Student records, and 150,000 Temporary work (skilled) records. This would undoubtedly provide for a much greater degree of detailed analysis than is currently possible with the ABS *Characteristics of Recent Migrants Survey* (CORMS) (ABS, 2013b).

Detailed analysis of labour force characteristics, such as labour force participation and unemployment rates would be possible, with more detailed analyses by age and sex. It may also be possible to analyse relatively small occupation groups (to 4 digit level). Characteristics of temporary migrants by location would be feasible for relatively small areas where larger groups of temporary migrants cluster (around universities, or in particular regions needing temporary migrant labour).

It is likely the creation of a future ACTMID file would produce a useful dataset for analysis, particularly for the largest groups of temporary migrants (International Students and Temporary work (skilled) subclass 457).

However, more work would need to be done to assess and improve data quality and support more detailed analysis (e.g. to enhance analysis of smaller visa subclass populations, or to do finer regional analysis, or other finer disaggregation).

REFERENCES

Australian Bureau of Statistics (1997) *ASCO – Australian Standard Classification of Occupations, Second Edition, 1997* , cat. no. 1220.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1220.0 >

—— (2011) *Standard Australian Classification of Countries (SACC), 2011*, cat. no. 1269.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1269.0 >

—— (2013a) *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 >

—— (2013b) *Characteristics of Recent Migrants, Australia*, cat. no. 6250.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/6250.0 >

—— (2014) *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 >

Chipperfield, J.O. and Chambers, R.L. (2015) "Using the Bootstrap to Analyse Binary Variables with Probabilistically-Linked Data", *Journal of Official Statistics*, Accepted for publication.

Department of Human Services (2014) *Eligibility for Medicare Card*, DHS website, DHS, Canberra.
< http://www.humanservices.gov.au/customer/enablers/medicare/medicare-card/eligibility-for-medicare-card >

Department of Immigration and Multicultural and Indigenous Affairs (2005) *The Global Systems Environment Program*, Department of Immigration and Border Protection website, DIBP, Canberra.
< http://www.immi.gov.au/about/speeches-pres/industry-briefings/_pdf/GSE.pdf >

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

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 >

Samuels, C. (2012) "Using the EM Algorithm to Estimate the Parameters of the Fellegi-Sunter Model for Data Linking", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.120, Australian Bureau of Statistics, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.120 >

Walsh, L. and Weckert, A. (2014) "Feasibility Study of Linking Migrant Settlements Records to Personal Income Tax Data", *Methodology Research Papers*, cat. no. 1351.0.55.051, Australian Bureau of Statistics, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.051 >

All URLs last viewed on 21 October 2014

GLOSSARY

Accuracy

When linking records from FileA to FileB, the Accuracy is the measure of how correctly the record linkage project has identified the true positives (matches) and true negatives (non-matches) from FileA.

Blocking

Blocking is a process of grouping data together before attempting to match records between two data files. The data is 'blocked' into groups and only the records within the group (e.g. overseas born; or adults aged over 15 years) are compared using the selected linking variables. The process is used to reduce the number of comparisons required to find a potentially matching record from each dataset. Most variables can be used for blocking, however, some variables work better than others. By changing the blocking variables for each pass, potential matches that have been missed can be picked up by using a different variable the next time.

Calibration

When producing a linked data set, the missing data that arises from unlinked records is dealt with through a process of calibration. This process uses information about unlinked records (known from the administrative sources) to adjust for missed links. This is somewhat different to a normal weighting process that is undertaken with survey samples that are weighted to a known population using population benchmarks.

Cut-off weight

The record weight pairs generated for probabilistic record linkage allow linked record pairs to be ranked according to the probability that they are indeed matches. Typically, the record linkage process proceeds by identifying the critical weights (cut-off weights) that partition the linked record pairs into probable matches, possible matches and probable non-matches. The simulations conducted in the Appendix are designed to assist with the identification of effective cut-off weights.

Link

A record pair which was brought together in the linking process. A link may be true or false (i.e. may not be a match of the same person).

Link Accuracy (Precision)

When linking records from FileA to FileB, the Link Accuracy or Precision is the proportion of all linked record pairs that are in fact true (correct) matches.

Link Rate

When linking records from FileA to FileB, the Link Rate is the proportion of FileA records that have been linked.

m-probability

The probability that two records will agree or disagree on a given characteristic, given that both records relate to the same person.

Match

A true record pair belonging to the same person.

Pass

A (linking) pass is characterised by a selection of blocking and linking variables used to compare and link records from two data files. Probabilistic record linkage usually involves multiple linking passes conducted sequentially, with each successive pass searching for additional matches that may have been missed by the preceding passes. Each pass will typically employ different blocking and linking variables or different comparison and decision criteria from the preceding passes.

Recall rate

When linking records from FileA to FileB, the Recall rate is the proportion of all true (correct) matches that are actually linked.

Run

A (linking) run is essentially a stand-alone linking pass (see "Pass" above). In the Appendix, the term "run" is preferred to "pass" to emphasise that the simulation modelling independently evaluates each selection of blocking and linking variables, and is not capable of modelling the cumulative outcomes from conducting multiple sequential passes.

u-probability

The probability that two records will agree or disagree on a given characteristic, given that the records relate to different individuals.

APPENDIX

A. AN ASSESSMENT OF THE FEASIBILITY OF LINKING TEMPORARY VISA HOLDER RECORDS TO THE 2011 CENSUS

This report summarises the results of a simulation study conducted to assess the feasibility of locating the Census records of temporary visa holders who were in Australia at the time of the 2011 Census.

A.1 The Temporary Visa Holders (TVH) and 2011 Census data files

The TVH data file contains records for 513,184 persons who had been granted temporary visas to undertake employment or study in Australia, and were presumed to be living in Australia on Sunday, 31 July 2011. This date represents the end of the reporting month which is closest to the date of the 2011 Australian Census of Population and Housing. The Census was conducted on Tuesday, 9 August 2011.

For an overview of the TVH data file, and a discussion of the available data items, please review Section 3 of this paper.

The Census data file extracted for the purposes of this study includes all respondents to the 2011 Census who did not explicitly identify themselves as Australian citizens on Census night. This group comprises about 2.5 million persons, many of whom are long-term residents.

It was decided that Census respondents who identified themselves as 'International Visitors' should be excluded from the data file. Potentially, some temporary visa holders may have incorrectly identified themselves with this category. If so, there is little or no prospect of locating their Census records, since detailed responses are not sought from international visitors.

A.2 Data items

A review of the information contained in the two data files identified the data items listed in table A.1 as being potentially comparable, and therefore informative for conducting record linkage.

Section 3 contains information on the derivation and quality of the key data items required for this study. However, a couple of data items require further explanation.

Date of birth and Age

Census respondents have the option of reporting either their exact date of birth or their age, and a significant proportion of Census records (10%) do not have exact date of birth information. Of those respondents who do not provide complete date of birth information, very few supply partial information (e.g. day only or month only).

A.1 Summary of the data items extracted from the Census and TVH data files

		Census	*Temporary Visa Holders*	
Data items extracted		*Data item*	*Data item*	*Modification*
MB	Meshblock	MBUCP	Address fields	Coded to MB
SA1	Statistical Area 1	SA1UCP	Address fields	Coded to SA1 (11 digit)
DAY	Day of birth	DOBDP	vh_birth_dd	
MONTH	Month of birth	DOBMP	vh_birth_mm	
BYEAR	Year of birth	DOBYP	vh_birth_yy	
AGE	Age	AGEP		Age computed from date of birth
SEX	Sex	SEXP	vh_gender_ds	Text converted to numeric
MST	Marital status	MSTP	vh_marital_status_ds	Text converted to numeric
COB	Country of birth	BPLP	vh_birth_cntry_cd	Text converted to SACC
CIT	Citizenship	CITP	vh_citz_cntry_cd	Text converted to SACC
YOA	Year of arrival	YARP	vh_grant_dt	Proxy for date of arrival

As it would be invalid to simulate independent responses to the day and month of birth fields, these fields have been combined to form a single field, representing the day of the year. Each of the 366 day and month combinations is assigned the same numerical code, regardless of leap years (i.e. 01JAN = 1, 01MAR = 61, 31DEC = 366).

Age (on Census night) – whether reported directly or computed from date of birth – is available for all Census records, and can be derived for virtually all TVH records.

Marital status

The marital status field extracted from the Census is focussed on recording whether the respondent is currently (or has previously been) in a registered marriage. The TVH field is more inclusive of *de facto* relationships, and does identify these separately.

For the purposes of this study, different categories were used for Marital status than the standard Census categories.

Acceptable comparability can be achieved by collapsing the responses found on both data files to the following definition:

$$\text{MST} = \begin{cases} 0 & : \text{Aged } 0-14 \text{ years;} \\ 1 & : \text{Never married (incl. } \textit{de facto}\text{);} \\ 2 & : \text{No longer in a registered marriage;} \\ 3 & : \text{Currently in a registered marriage.} \end{cases}$$

It is true that several TVH applicants aged under 15 years had a recorded marital status other than 'never married'. This may be true also for Census respondents, but this information has been lost as all persons under 15 are assigned the marital status of 'not applicable'.

For record linkage purposes, the marital status field provides additional information to distinguish between persons aged 15 years and over, but provides no additional information about children. In fact, including both age and marital status as linking fields should be avoided as this will tend to distort the aggregate record pair weights for children.

With this in mind, it is useful to define the closely related data item MSTA, which is identical to MST except that persons aged under 15 years are coded to "missing" rather than "0". By comparing MSTA on the TVH data file with MST on the Census file, TVH applicants aged 15 years and over can be distinguished from Census respondents aged 15 years and over on the basis of marital status, without creating any distortions arising from the simultaneous use of age as a linking field.

A.3 Blocking and linking strategies

The data items identified in table A.1 are typical of the data items employed in previous Bronze standard probabilistic record linkage projects involving Census data. Two features are particularly relevant in this study:

- Address information, which is typically highly influential in the record linkage process, is unavailable for a significant proportion of TVH records (negative), and
- Country of birth / year of arrival information will be much more informative in this study than in projects that do not specifically focus on migrant populations (positive).

Five blocking and linking strategies (labelled RUN 1 – RUN 5) have been devised for this study. They are summarised in table A.2.

By simulating the likely outcomes from these five runs, it is hoped to provide relevant insights into the overall potential of the proposed record linkage project. The runs themselves are not intended to be prescriptive of the runs that may be undertaken should the proposed project proceed.

A fundamental assumption underlying the design of these blocking and linking strategies is that they will be used for probabilistic record linkage using the methodological framework devised by Fellegi and Sunter (1969).

Although there are relatively few data items to select from, it is important to avoid undue reliance upon key data items across all strategies, when combinations of other data items may prove equally capable of locating matches.

A.2 Blocking and linking strategies

	RUN 1	*RUN 2*	*RUN 3*	*RUN 4*	*RUN 5*
MB	MB	—	—	—	—
SA1	—	SA1	SA1	—	—
BDAY	BDAY	BDAY	BDAY	BDAY	BDAY
BYEAR	—	BYEAR	—	—	—
AGE	AGE	—	AGE (±1)	AGE	AGE
SEX	—	SEX	SEX	SEX	SEX
MST	MSTA	MSTA	MSTA	MSTA	MSTA
COB4	COB4	COB4	—	COB4|COB2	COB4
COB2	—	—	COB2|COB1	COB2	—
COB1	—	—	COB1	—	—
YOA	YOA	YOA	YOA (±1)	YOA (±1)	YOA

Notes:
Shading indicates blocking fields.
(1)|(2) indicates that the significance of agreement on the first linking field is to be assessed conditional upon the knowledge that agreement on the second field had already been observed..
(±1) indicates that a tolerance of ± one (year) is permitted when testing for agreement.

RUN 1 is designed to locate the most definite matches in the study, as it applies the strictest tests of agreement to all fields – most detailed geography (MB), exact agreement on date of birth / age, marital status and year of arrival, and most precise identification of country of birth (four-digit). Chance agreement on five or more of the six fields would be considered extremely improbable (except for identical twins), thus allowing for some missingness or disagreement (as a result of reporting error) to be overlooked. This may prove convenient, for example, for matching visa holders who did not supply their date of birth in the Census.

Sex, which typically has a very high penalty for disagreement, is not included among the linking fields for RUN 1 as it is unlikely to be required for confirmation of clear matches – although it could perhaps be reintroduced to resolve poorer quality links.

All other blocking and linking strategies are designed to relax the strictness of the tests in RUN 1, in a way which recognises the real potential for individuals to be reported inconsistently on the two data files. Typical inconsistencies will arise when the two records do not relate to the same point in time (e.g. changing addresses) or where data items on the two data files are not as comparable as presumed (e.g. Census year of arrival and year of TVH approval).

RUN 2, requiring agreement on exact date of birth and sex, is not strictly constrained by the condition that potential matches *must* agree on geography. Nonetheless, agreement on SA1 still provides strong confirmation, and disagreement is penalised. For the many TVH records that are *missing* this data item, though, agreement on the additional linking fields may be sufficient to confirm match status.

Like RUN 1, RUN 3 requires that potential matches agree on geography – in this case, the broader SA1 geography. A lesser degree of precision is also accepted for other key linking fields (age, country of birth and year of arrival). However, provided agreement is observed on most of the linking fields, the majority of matches (true positives) are still expected to be easily identified.

To recap, RUN 1 – RUN 3 should prove sufficient to identify matching records which agree on geography, or have missing geography – provided a sufficient level of agreement is observed on the supplementary linking fields. It remains to evaluate whether or not it may be possible to locate matches that *disagree* on geography. RUN 4 and RUN 5 are proposed to investigate this question.

The geographical data items (MB and SA1) are not used as either blocking or linking fields in RUN 4 and RUN 5. To compensate, agreement on country of birth is now a pre-requisite for identifying matches. Matches are not expected to be identifiable without strict agreement on birthday and age, but these data items are left as linking fields to test this assumption.

Simulation of the record linkage process defined by the above blocking and linking strategies requires the following inputs:

- estimates of the parameters required by the Fellegi-Sunter methodology – especially the m- and u-probabilities pertaining to the specified data items; and
- detailed information about the properties of the respective blocking fields.

Some of these inputs may be obtained by extracting factual information from the two data files, but there is also a need to employ some assumptions (either as informed judgement or as hypothetical scenarios).

The following two sections describe the collation of the model inputs and the running of the simulation models.

A.4 Estimating the parameters

The key parameters required for the simulation process are the m- and u-probabilities that underpin the Fellegi-Sunter model. These may be derived by combining available factual information from the two data files with hypothetical assumptions about the comparability of data items for matching record pairs.

A.4.1 The facts

Simple interrogation of the TVH and Census data files can provide frequency counts of the unique responses for all data items of interest on each data file.

A.3 Global counts of agreement, disagreement and missingness

	Agree	*Disagree*	*Missing*	*All comparisons*
MB	17,079,606	846,642,398,154	424,677,417,824	1,271,336,895,584
SA1	51,065,850	850,002,382,780	421,283,446,954	1,271,336,895,584
BDAY	3,151,166,256	1,138,981,323,820	129,204,405,508	1,271,336,895,584
BYEAR	21,862,077,740	1,121,501,610,580	127,973,207,264	1,271,336,895,584
AGE	24,351,631,314	1,246,893,602,283	91,661,987	1,271,336,895,584
AGE (±1)	72,804,816,694	1,198,440,416,903	91,661,987	1,271,336,895,584
SEX	635,688,744,374	635,628,332,402	19,818,808	1,271,336,895,584
MSTA—MST	376,529,449,199	740,111,740,541	154,695,705,844	1,271,336,895,584
COB4	38,322,960,830	1,130,516,086,224	102,497,848,530	1,271,336,895,584
COB2	83,960,522,770	1,087,835,946,268	99,540,426,546	1,271,336,895,584
COB1	134,657,618,077	1,037,138,850,961	99,540,426,546	1,271,336,895,584
COB4|COB2	38,322,960,830	45,340,554,633	296,995,194	83,960,522,770
COB2|COB1	83,960,522,770	50,697,090,175	0	134,657,618,077
YOA	91,162,658,398	867,524,582,946	312,649,654,240	1,271,336,895,584
YOA (±1)	252,902,371,620	705,784,869,724	312,649,654,240	1,271,336,895,584
MB	17,079,606	846,642,398,154	424,677,417,824	1,271,336,895,584
BDAY & BYEAR & SEX	30,647,023	1,140,778,095,857	130,528,152,704	1,271,336,895,584
SA1 & COB1 & SEX	7,902,021	783,485,320,127	487,843,673,436	1,271,336,895,584
AGE & COB2 & SEX	1,471,019,314	1,170,222,696,814	99,643,179,456	1,271,336,895,584
COB4 & SEX	19,459,561,453	1,149,365,769,355	102,511,564,776	1,271,336,895,584

Consider the global comparison space, in which every one of the 513,184 unique TVH records is compared with every one of the 2,477,351 Census records (a total of 1.27×10^{12} record pair comparisons).

From the frequency counts for each data item, it is straightforward to compute the numbers of pairwise comparisons that would be classified as agreement, disagreement or missing in the global comparison space. The results of these computations are shown in table A.3.

In simulating record linkage, it is important to recognise the role of missing responses. The consequences of a simulated missing response on a TVH record will differ from a simulated missing response on a Census record, and it is important to use the appropriate probabilities of missingness for each data file.

The frequencies of missing responses were obtained in the process of computing table A.3, and are presented now as proportions in table A.4. The first column reports the proportion of TVH records with a missing response, the second column reports the proportion of Census records with a missing response and the third column reports the number of record comparisons in the global comparison space in which one or both records have a missing response.

A.4 Proportion of missing responses

	Exact			*Rounded*		
	TVH	*Census*	*Combined*	*TVH*	*Census*	*Combined*
MB	0.33404003	0.00000000	0.33404003	0.333	0.000	0.333
SA1	0.33137042	0.00000000	0.33137042	0.333	0.000	0.333
BDAY	0.00007015	0.10156575	0.10162877	0.000	0.100	0.100
BYEAR	0.00000000	0.10066034	0.10066034	0.000	0.100	0.100
AGE	0.00007210	0.00000000	0.00007210	0.000	0.000	0.000
AGE (±1)	0.00007210	0.00000000	0.00007210	0.000	0.000	0.000
SEX	0.00001559	0.00000000	0.00001559	0.000	0.000	0.000
MSTA—MST	0.12167955	0.00000000	0.12167955	0.122	0.000	0.122
COB4	0.00368289	0.07722362	0.08062210	0.004	0.077	0.081
COB2	0.00000585	0.07829048	0.07829587	0.000	0.078	0.078
COB1	0.00000585	0.07829048	0.07829587	0.000	0.078	0.078
COB4|COB2	0.00367707	0.00000000	0.00367707	0.004	0.000	0.004
COB2|COB1	0.00000000	0.00000000	0.00000000	0.000	0.000	0.000
YOA	0.00000000	0.24592195	0.24592195	0.000	0.246	0.246
YOA (±1)	0.00000000	0.24592195	0.24592195	0.000	0.246	0.246
MB	0.33404003	0.00000000	0.33404003	0.333	0.000	0.333
BDAY & BYEAR & SEX	0.00008574	0.10259305	0.10266999	0.000	0.100	0.100
SA1 & COB1 & SEX	0.33137822	0.07829048	0.38372494	0.333	0.078	0.385
AGE & COB2 & SEX	0.00009353	0.07829048	0.07837669	0.000	0.078	0.078
COB4 & SEX	0.00369458	0.07722362	0.08063289	0.004	0.077	0.081

In the simulation process, it is assumed that the probability of observing a missing response on one data file is independent of the probability of observing a missing response on the other, and this is used to compute the combined proportions.

The rounded proportions in table A.4 are a simplified representation that will be used in the next section to derive consistent m- and u-probabilities.

A.4.2 The assumptions

Section A.4.1 summarised factual information for the global comparison space. In this section a series of assumptions will be used to characterise the two disjoint sets which together comprise the global comparison space – the set of all matching record pairs (M) and the set of all non-matching pairs (U).

The first assumption is that every TVH record has a matching Census record. This is almost certainly false, but as there is no possibility of quantifying the real proportion, it provides a sensible hypothetical starting point. The assumption does not imply that every TVH record *can* be matched, and the reality is that most matches that fail will fail because they do not agree on any choice of blocking fields. This latter condition is generally easier to quantify.

A.5 Derivation of m- and u-probabilities from assumptions

		m-probabilities			*u-probabilities*		
	Assumptions	*agree*	*disagree*	*missing*	*agree*	*disagree*	*missing*
MB	0.900	0.600	0.067	0.333	0.000013	0.666987	0.333
SA1	0.920	0.615	0.052	0.333	0.000040	0.666960	0.333
BDAY	0.975	0.875	0.025	0.100	0.002478	0.897522	0.100
BYEAR	0.960	0.863	0.037	0.100	0.017196	0.882804	0.100
AGE	0.960	0.960	0.040	0.000	0.019154	0.980846	0.000
AGE (±1)	0.980	0.980	0.020	0.000	0.057266	0.942734	0.000
SEX	0.999	0.999	0.001	0.000	0.500000	0.500000	0.000
MSTA—MST	0.940	0.826	0.052	0.122	0.296168	0.581832	0.122
COB4	0.880	0.809	0.110	0.081	0.030144	0.889164	0.081
COB2	0.900	0.830	0.092	0.078	0.066041	0.855959	0.078
COB1	0.950	0.876	0.046	0.078	0.105918	0.816082	0.078
COB4|COB2	0.980	0.976	0.020	0.004	0.456438	0.543558	0.004
COB2|COB1	0.990	0.990	0.010	0.000	0.623510	0.376490	0.000
YOA	0.850	0.640	0.114	0.246	0.071706	0.682294	0.246
YOA (±1)	0.930	0.700	0.054	0.246	0.198926	0.555074	0.246
MB	0.900	0.600	0.067	0.333	0.000013	0.666987	0.333
BDAY & BYEAR & SEX	0.960	0.860	0.040	0.100	0.000024	0.899976	0.100
SA1 & COB1 & SEX	0.875	0.540	0.075	0.385	0.000006	0.614968	0.385
AGE & COB2 & SEX	0.860	0.792	0.130	0.078	0.001157	0.920843	0.078
COB4 & SEX	0.880	0.809	0.110	0.081	0.015306	0.904002	0.081

Note: The assumptions shown in column 1 specify the proportion of agreement expected from non-missing comparisons of matching records.

The second assumption is that missingness is uninformative. Expressed another way, the probability of observing missingness for a given field comparison is identical for all record pairs, regardless of whether the pair is a match or a non-match. That is, the m- and u-probabilities for missingness are identical, resulting always in zero field weights for missingness.

This assumption provides the starting point for populating table A.5. The global probabilities of missingness were reported in table A.4, and are now entered into both the m-probabilities and u-probabilities columns for missing.

In almost every case, there is no significant difference between the u-probabilities of agreement and disagreement and the observed global probabilities – since the only difference arises from subtracting the relatively insignificant contribution of the matching records.

To derive m-probabilities, though, it is necessary to employ additional assumptions. Assumption 1 established that the number of matching pairs is the same as the number of TVH records. Assumption 2 established the numbers of missing and non-missing comparisons for each field. Now it is necessary to introduce assumptions

regarding the split between agreement and disagreement for the non-missing comparisons. Typically this split will be different for every linking field. For example, matching records will almost always agree on sex, but for valid reasons may not always agree on geography or marital status (as these responses can change over time).

Table A.5 provides (without explanation at this time) a set of assumptions regarding the proportion of non-missing responses that are expected to agree on matching records. These assumptions are then applied to generate the m-probabilities for agreement and disagreement.

For example, 0.333 of matching records are assumed to have a missing comparison on Meshblock. Of the non-missing comparisons, 0.900 are assumed to agree, leading to an m-probability of agreement of $0.900 \times (1 - 0.333) = 0.600$.

It is now possible to estimate actual counts for the numbers of matching records that display agreement, disagreement and missingness – by distributing the total count of TVH records (513,184) in proportion to the assumed m-probabilities. Subtracting these counts from the global counts recorded in table A.3 provides estimates of the corresponding counts for the set of non-matching comparisons (U). u-probabilities for agreement, disagreement and missingness may now be deduced.

A.4.3 The parameters

The previous sections (A.4.1 and A.4.2) have described the logic used to combine factual information with clearly documented assumptions to develop estimates of the key parameters required to conduct simulation studies of the blocking and linking strategies proposed in Section A.3.

Note that the methodology employed to derive the parameter estimates utilises only summary statistics acquired independently from the two data files. This acknowledges a situation that is typically encountered in feasibility analysis – the analyst does not have simultaneous access to both data files, and therefore cannot utilise statistical methods such as the EM Algorithm (see, for example, Samuels, 2012) to estimate the m- and u-probabilities.

The essential simulation parameters have been extracted from tables A.4 and A.5, and summarised in table A.6. From these selected parameters it is possible to reconstruct the full set of m- and u-probabilities required for all blocking and linking strategies.

Table A.7 and figure A.8 report the field weights that correspond to the m- and u-probabilities shown in table A.5 / table A.6.

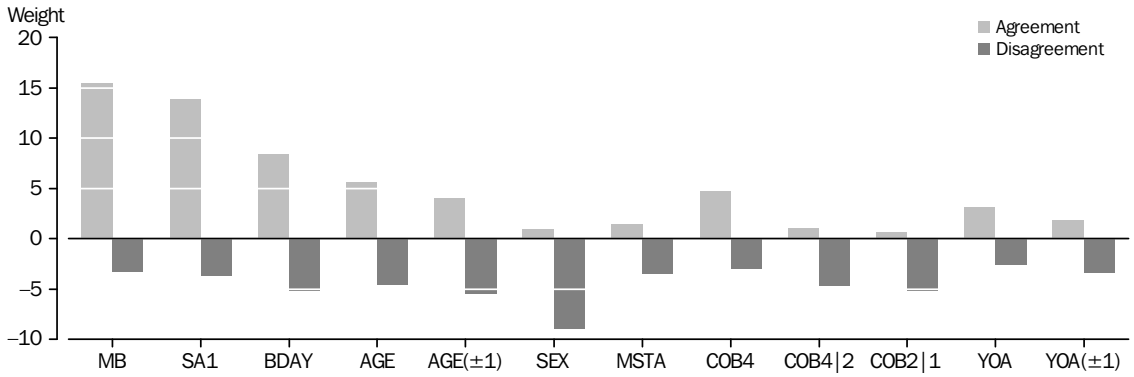
A.6 Input parameters for the simulation models

	m-agree	*u-agree*	*missing-a*	*missing-b*
MB	0.600	0.000013	0.333	0.000
SA1	0.615	0.000040	0.333	0.000
BDAY	0.875	0.002478	0.000	0.100
BYEAR	0.863	0.017196	0.000	0.100
AGE	0.960	0.019154	0.000	0.000
AGE (±1)	0.980	0.057266	0.000	0.000
SEX	0.999	0.500000	0.000	0.000
MSTA—MST	0.826	0.296168	0.122	0.000
COB4	0.809	0.030144	0.004	0.077
COB2	0.830	0.066041	0.000	0.078
COB1	0.876	0.105918	0.000	0.078
COB4|COB2	0.976	0.456438	0.004	0.000
COB2|COB1	0.990	0.623510	0.000	0.000
YOA	0.640	0.071706	0.000	0.246
YOA (±1)	0.700	0.198926	0.000	0.246

A.7 Global field weights and implicit blocking weights

	agree	*disagree*	*missing*
MB	15.494163	–3.315426	0.000000
SA1	13.908299	–3.681017	0.000000
BDAY	8.463963	–5.165947	0.000000
BYEAR	5.649216	–4.576496	0.000000
AGE	5.647317	–4.615955	.
AGE (±1)	4.097031	–5.558779	.
SEX	0.998557	–8.965784	.
MSTA—MST	1.479726	–3.484019	0.000000
COB4	4.746197	–3.010912	0.000000
COB2	3.651677	–3.217836	0.000000
COB1	3.047983	–4.149008	0.000000
COB4|COB2	1.096462	–4.753717	0.000000
COB2|COB1	0.667016	–5.234540	.
YOA	3.157906	–2.581360	0.000000
YOA (±1)	1.815123	–3.361649	0.000000
MB	15.494163	–3.315426	0.000000
BDAY & BYEAR & SEX	15.129015	–4.491815	0.000000
SA1 & COB1 & SEX	16.457637	–3.036049	0.000000
AGE & COB2 & SEX	9.418968	–2.824444	0.000000
COB4 & SEX	5.723970	–3.034789	0.000000

A.8 Global field weights



The lower sections of tables A.5 and A.7 report the probabilities and weights that correspond to the combinations of blocking fields proposed for RUN 1 – RUN 5. Two aspects are particularly relevant:

1. The assumptions documented in table A.5 will be used in Section A.5 to construct the input data files that define the block sizes for simulation. Specifically, they will be used to estimate the proportion of TVH records that are located in the same block as their matching Census record.

2. The agreement weights in table A.7 quantify the implicit contribution to record pair weights arising from agreement on the blocking fields. These implicit weights can be added to the record pair weights generated within their respective runs to produce comparable weights across all blocking and linking strategies.

A.5 The simulations

The parameters computed in Section A.4 can be used to simulate patterns of agreement / disagreement / missing on selected linking fields, conditional upon whether the simulated record pair is assumed to be a matching pair or a non-matching pair.

Conceptually, each TVH record will be paired with all Census records that agree on the same selection of blocking fields. One of those Census records *may* be the matching record that is sought. The simulation model simulates the agreement patterns for all candidate matches to the TVH record, and then chooses the record pair with the highest record pair weight (or selects randomly from all pairs which share the highest record pair weight). The simulation model then records how frequently the selected record pair is indeed the correct match for the TVH record.

To generate useful predictions, the simulation model requires knowledge of the block characteristics for the selected blocking strategies, and also information on the expected proportion of matching pairs that will agree on the blocking fields. This information is supplied to the model in an input data file.

A.5.1 The input data file

Table A.9 illustrates the information required for the input data file. Simple interrogation of the data files will produce the numbers of TVH and Census records per block. However, the final column of data – the expected number of matching pairs within each block – presents difficulties. Clearly this information cannot be known, or even estimated with any certainty, and further assumptions must be employed.

A.9 Structure of the input data file

Block no.	*No. of TVH records in block*	*No. of Census records in block*	*No. of TVH records that share the same block as their matching Census record*
1	n_1^{TVH}	n_1^{Census}	n_1^{Match} $(\leq n_1^{\text{TVH}})$
⋮	⋮	⋮	⋮
K	n_K^{TVH}	n_K^{Census}	n_K^{Match} $(\leq n_K^{\text{TVH}})$

Having obtained counts of the numbers of TVH and Census records per block, two initial consistency checks are required:

- If $n_k^{\text{TVH}} = 0$ or $n_k^{\text{Census}} = 0$, then drop block k from the simulation; and
- If $n_l^{\text{TVH}} > n_l^{\text{Census}}$ then set $n_l^{\text{TVH}} = n_l^{\text{Census}}$.

Table A.10 reports, for all five blocking strategies, the total number of blocks remaining, and the aggregate numbers of TVH and Census records remaining following application of the consistency checks. Also reported is the total number of record pair comparisons that will be investigated / simulated under each blocking strategy.

In Section A.4.2, estimated m-probabilities of agreement were developed for the combinations of blocking fields selected for RUN 1 – RUN 5. These m-probabilities took into account the assumptions and observed missingness for the relevant individual data items. From these probability estimates, estimates of the aggregate numbers of matching TVH and Census records were calculated, and these estimates are supplied also in table A.10.

To estimate the n_k^{Match} counts for individual blocks, k, a process of random selection was employed. For example, for RUN 1, 307,584 records were randomly selected from the 328,739 records that remained within scope following the consistency checks. The distribution of the randomly selected records across the blocks was then used to supply the required estimates.

A.10 Blocking strategies – overview

	RUN 1	*RUN 2*	*RUN 3*
Number of blocks	86,424	41,077	112,960
Number of TVH records	328,739	511,692	305,082
Number of Census records	1,388,003	1,796,620	853,828
Number of matching records (est.)	307,584	442,075	276,730
Number of comparisons	16,938,387	30,582,694	7,794,471
Average comparisons / block	196	745	69
Median TVH records / block	2	6	1
Median Census records / block	12	39	4
Median comparisons / block	22	200	6

	RUN 4		*RUN 5*	
	Complete	*Reduced*	*Complete*	*Reduced*
Number of blocks	3,201	2,908	430	376
Number of TVH records	488,212	289,119	478,649	72,164
Number of Census records	1,898,110	1,467,043	2,272,956	388,277
Number of matching records (est.)	406,747	240,701	415,188	62,654
Number of comparisons	1,378,920,023	380,390,018	19,434,293,311	243,608,381
Average comparisons / block	430,778	130,808	45,196,031	647,895
Median TVH records / block	21	18	47	25
Median Census records / block	197	168	304	188
Median comparisons / block	3,312	2,875	14,262	5,353

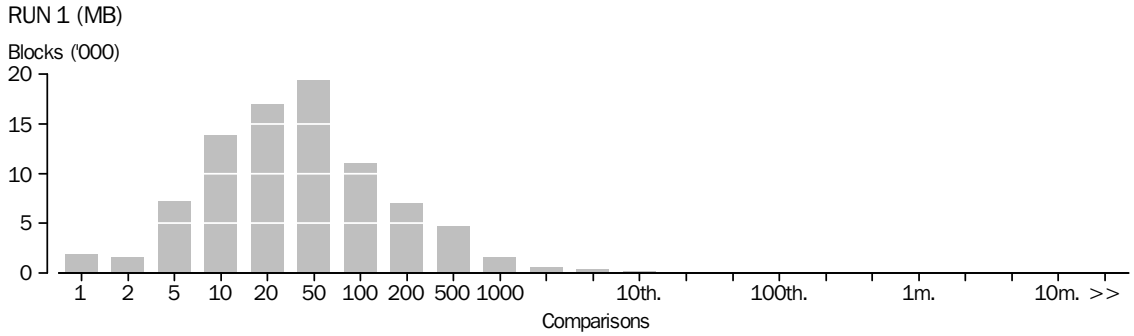
A.5.2 Other considerations

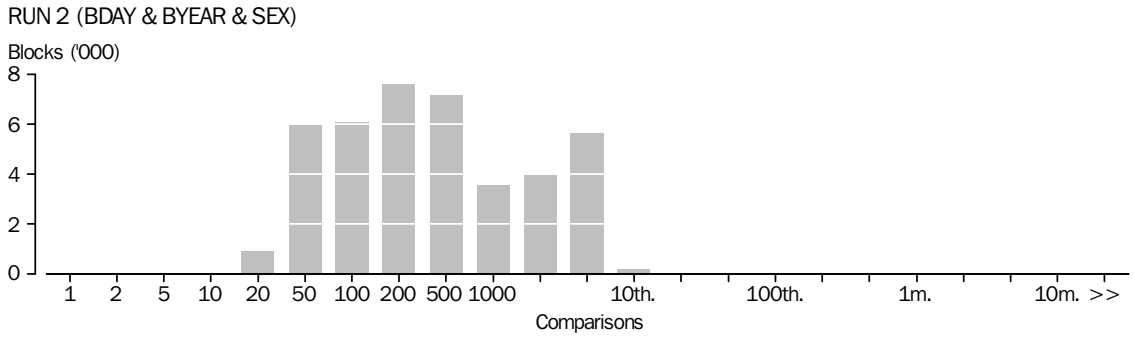
While table A.10 provides aggregate statistics on the five blocking strategies, figure A.11 highlights the wide disparity in block characteristics that can be observed within and across all blocking strategies. Note in particular the logarithmic scale used on the horizontal axes, and the scale factors used on the vertical axes.

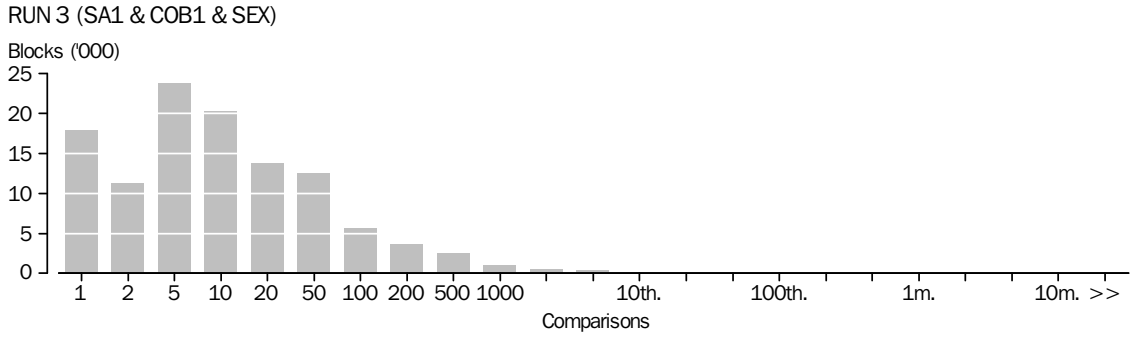
The distributions illustrated in figure A.11 represent factual information, and these facts contribute significantly to the simulation process. Matching records are more likely to be accurately identified within small blocks (e.g. one TVH record and five Census records) than large blocks (e.g. 1000 TVH records and 5000 Census records), and the simulation results will reflect this. It will be important to remember this qualification when interpreting the final model diagnostics – and acknowledge that the visa holders who fall within the smaller blocks of the blocking strategy will be matched more easily and accurately than those falling within the large blocks.

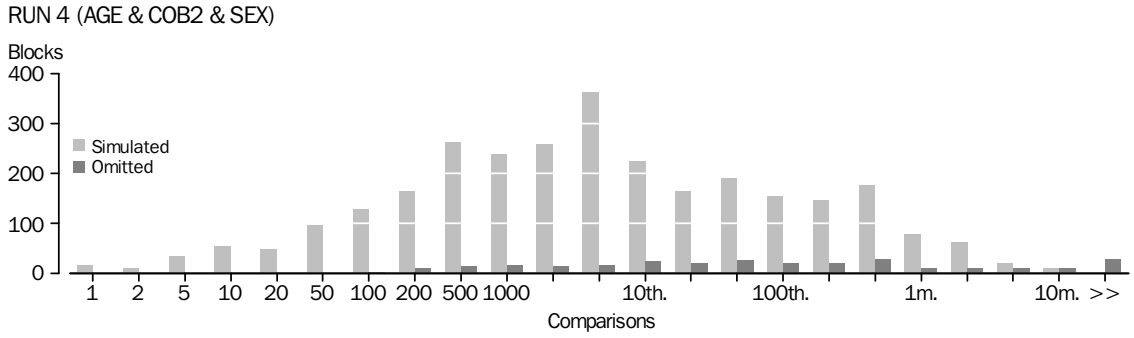
These observations are particularly pertinent to the evaluation of RUN 4 and RUN 5.

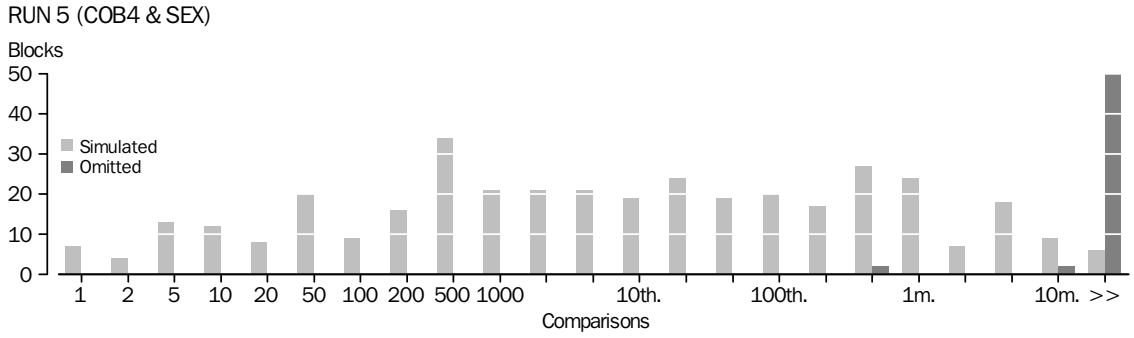
A.11 Block sizes











In Section A.3, the blocking strategies for RUN 4 and RUN 5 were justified on the basis that opportunity must be provided to locate visa holders who had changed or misreported their address – and would therefore be excluded or penalised heavily in the earlier runs. Consequently, geographical fields were dropped for RUN 4 and RUN 5, and country of birth was utilised for blocking.

In the first stage of testing, it was apparent that the blocking strategies for these runs were producing an extremely wide range of block sizes, with many blocks proving to be impractically large (e.g. greater than ten or twenty million comparisons per block). It was decided, therefore, to omit blocks corresponding to COB2=61 (Chinese Asia) and COB=71 (Southern Asia, including India) from the simulation of RUN 4, and to omit the 25 countries (identified by COB4) that generated the largest block sizes from RUN 5.

The decision to omit these countries effectively presumes that visa holders from those countries cannot be matched if they do not supply accurate address information. The objective for RUN 4 and RUN 5 is then to assess whether it is possible to match visa holders from countries that are less commonly represented in the migrant population.

A.6 The results

A.6.1 The diagnostics

Figure A.12 displays the diagnostic statistics generated by the five simulation runs, and table A.13 provides the corresponding numerical information at the key weight cut-offs.

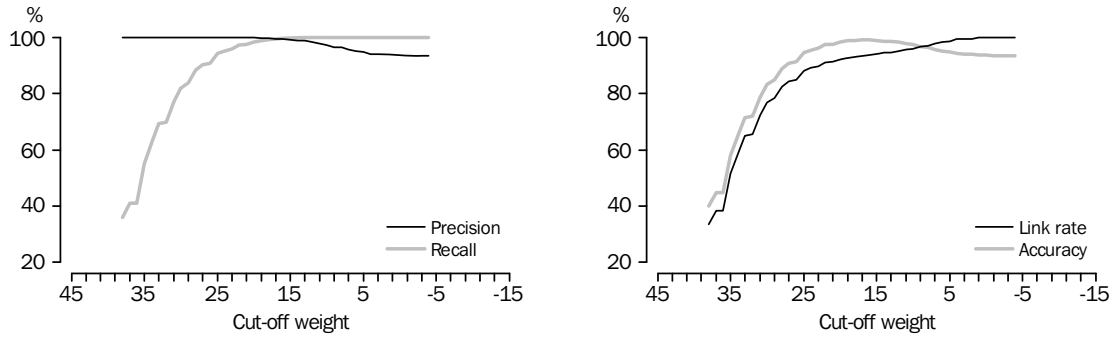
The diagnostics selected for presentation are the *precision* (or *link accuracy*), the *recall rate*, the *link rate* and the *accuracy* at integer cut-off weights. All weights include an adjustment for the implicit contribution from agreement on the respective blocking fields (as described in Section A.4.3), and therefore are comparable across the different runs.

The trade-off between precision and link rate is usually the primary focus in a feasibility study. It is important to note that the reported link rate tracks the proportion of records linked relative to the total number of records included in the blocking strategy – not the total number of TVH records.

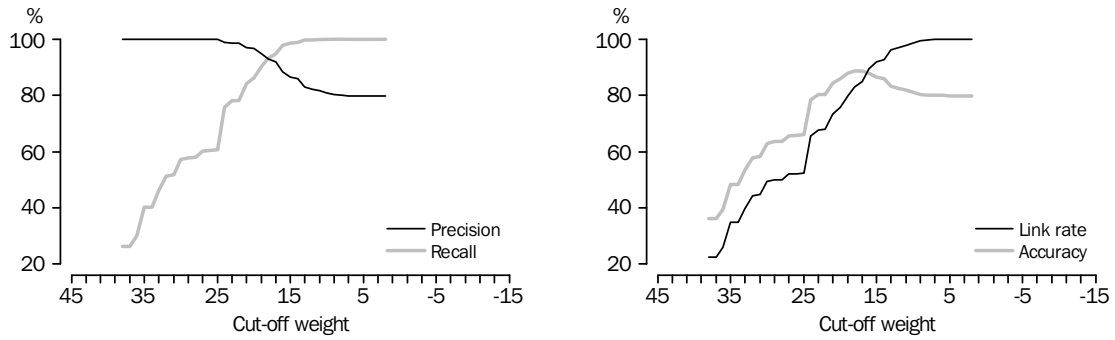
While not the primary focus, the recall rate and accuracy measures are useful in identifying a useful minimum threshold cut-off weight for the linkage. The point of intersection of the precision and recall rate curves identifies the cut-off at which the probability of misclassifying a match is equal to the probability of misclassifying a non-match. To lower the cut-off further would entail adding false links at a faster rate than true matches, and would not be recommended.

A.12 Simulation model diagnostics

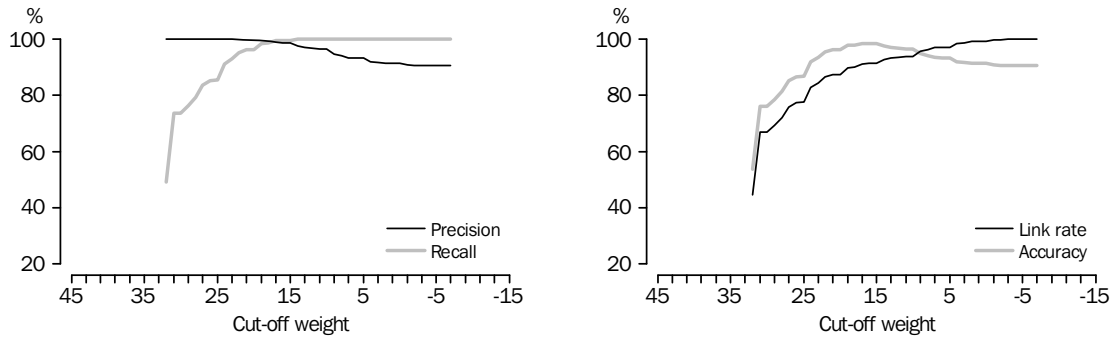
(a) RUN 1 (MB)



(b) RUN 2 (BDAY & BYEAR & SEX)



(c) RUN 3 (SA1 & COB1 & SEX)

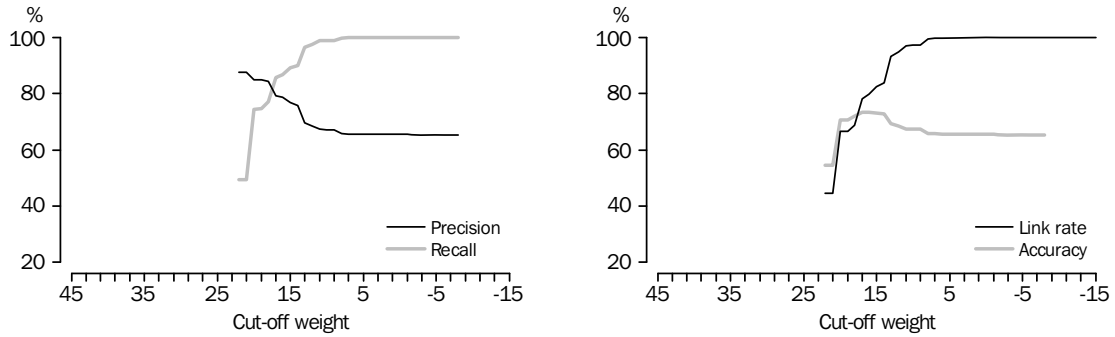


The accuracy statistic considers the correct classification of both matches and non-matches, and therefore is more comprehensive than the precision measure. Accuracy is maximised at the point where the best attainable classification of true positives and true negatives is observed.

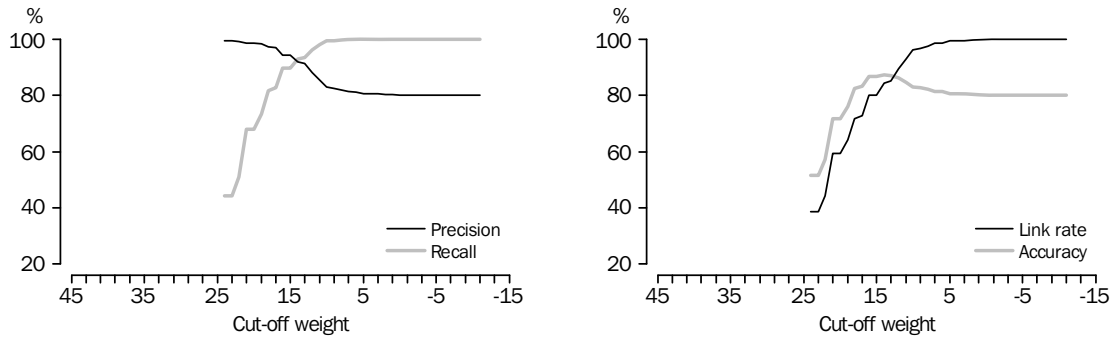
In most cases, the equal error rate cut-off will coincide closely with the point of maximum accuracy. Of course both points may identify a cut-off that produces an unacceptably low level of precision, or an insufficiently high link rate, and therefore it is inappropriate to identify such a cut-off as 'ideal'.

A.12 Simulation model diagnostics — continued

(d) RUN 4 (AGE & COB2 & SEX)



(e) RUN 5 (COB4 & SEX)



A.6.2 Commentary

At the point of maximum accuracy in RUN 1, it is predicted that 307,402 records can be linked, with a precision of 99.5%. RUN 3 is also predicted to have a very high precision (98.7%) at the point of maximum accuracy, but only 277,705 records are linked. As RUN 3 is designed to relax the strict comparison conditions of RUN 1, it is likely that the majority of records linked in RUN 3 will have also been linked in RUN 1, but it is not within the scope of this study to estimate the overlap.

Maximum accuracy in RUN 2 is observed at the same cut-off as RUN 1 and RUN 3, but the precision at that cut-off is probably unacceptably low (92.1%). At the cut-off for 99.0% precision, all 271,031 links feature agreement on SA1, and therefore almost certainly overlap with links made in RUN 1 and RUN 3. At the cut-off for 98.0% precision, an extra 76,125 records are linked, and the majority of these records have a missing response for SA1 – i.e. they could not have been linked in RUN 1 or RUN 3.

Simulation of RUN 4 provided no evidence of likely success. Even with complete agreement on all linking fields, precision is only predicted to be 87.5%. This is almost certainly due to the real probability of observing both true and false links with the same patterns of agreement on linking fields within the very large blocks.

A.13 Key threshold weights

			Linked		Not linked	
	Weight	*Adjusted weight*	*True Positive*	*False Positive*	*False Negative*	*True Negative*
RUN 1						
Perfect agreement (100.0%)	23.5	39.0	110,560	2	197,023	21,154
99.9% precision	5.6	21.1	299,699	266	7,778	20,996
Maximum accuracy (99.5%)	1.5	17.0	305,749	1,653	1,528	19,809
99.0% precision	–0.8	14.6	306,390	3,028	805	18,516
98.0% precision	–3.6	11.9	306,885	6,234	232	15,388
95.0% precision	–8.7	6.8	307,033	16,012	15	5,679
100% link rate (93.4%)	–18.9	–3.4	307,039	21,700	0	0
RUN 2						
Perfect agreement (100.0%)	23.3	38.4	115,851	0	326,224	69,617
99.9% precision	9.8	24.9	270,949	82	171,086	69,575
99.0% precision	9.8	24.9	270,949	82	171,086	69,575
98.0% precision	7.4	22.5	342,254	4,902	95,813	68,723
95.0% precision	4.6	19.8	381,914	18,672	47,261	63,845
Maximum accuracy (92.1%)	2.5	17.7	397,983	34,018	24,013	55,678
100% link rate (79.8%)	–12.8	2.4	408,180	103,512	0	0
RUN 3						
Perfect agreement (99.98%)	16.5	33.0	135,734	32	140,967	28,349
99.9% precision	8.8	25.3	236,239	108	40,413	28,322
99.0% precision	1.4	17.9	272,487	2,663	3,532	26,400
Maximum accuracy (98.7%)	0.7	17.1	274,118	3,587	1,823	25,554
98.0% precision	–1.6	14.9	274,694	5,023	1,174	24,191
95.0% precision	–6.6	9.9	275,542	11,290	201	18,049
100% link rate (90.4%)	–22.8	–6.3	275,661	29,421	0	0
RUN 4						
Perfect agreement (87.5%)	12.9	22.3	112,404	16,117	115,724	44,874
Maximum accuracy (78.6%)	6.8	16.2	180,772	49,307	27,566	31,474
100% link rate (65.4%)	–16.8	–7.3	188,930	100,189	0	0
RUN 5						
Perfect agreement (99.3%)	18.7	24.5	27,701	186	34,795	9,482
99.0% precision	17.3	23.0	31,786	255	30,653	9,470
98.0% precision	13.1	18.8	45,566	818	16,484	9,296
95.0% precision	10.6	16.4	51,664	1,872	9,774	8,854
Maximum accuracy (91.7%)	8.5	14.2	55,776	5,027	4,256	7,105
100% link rate (80.0%)	–15.8	–10.1	57,745	14,419	0	0

Notes:
1. Adjusted weights include the implicit contribution from agreement on the blocking fields.
2. Numbers in brackets report the estimated precision at the corresponding weight cut-off.

Inspection of the detailed results for RUN 4 emphasises the critical importance of observing agreement on BDAY, suggesting that an alternative strategy might be considered with BDAY added to the blocking fields (or perhaps replacing AGE).

There are also many alternative strategies that might be investigated for excluding large blocks from the linkage. As suggested in Section A.5.2, it is quite likely that records in the smaller blocks of RUN 4 have been linked with much greater precision than the overall diagnostics suggest.

RUN 5 utilises almost exactly the same data fields as RUN 4, and shares the same problem that matches are difficult to confirm without address information. Perhaps because the reduced RUN 5 is more closely targeted to finding visa holders from countries that have a lower representation in the migrant community, the diagnostics suggest that a reasonable proportion of the remaining TVH records may be linked with relatively high precision. For example, about 75% of the selected TVH records (i.e. about 53,500 records) can be linked at greater than 96% precision. It is beyond the scope of this study, however, to estimate how many of these linked records have inconsistent address information (and are therefore uniquely located by RUN 4 and RUN 5).

A.7 Conclusion

The simulation results for RUN 2 suggest that 347,156 records (67.6% of all TVH records) can be linked with 98% precision. In fact, those matching record pairs that agree on MB or SA1 can probably be located with greater precision in either RUN 1 or RUN 3.

RUN 1 and RUN 3 will also contribute extra matches for the record pairs that do not meet the blocking criteria of RUN 2. In particular, RUN 2 excludes the 10% of Census records that did not report exact date of birth.

The simulation of RUN 5 demonstrates some limited potential to match records that disagree on address information, but it is not possible to speculate how many such records there are. RUN 4 might also be revised or modified to assist in this search, but the present simulations do not provide convincing evidence.

Taking into account the probable overlap of the five independent blocking and linking strategies, and subject to the validity of all the documented assumptions, the following feasibility assessment is put forward:

> There is potential to link 70% of the records on the Temporary Visa Holders data file to the 2011 Census with a precision (or link accuracy) of 98% or higher.

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