General and Methodological Issues of the Census Data Enhancement Project

Glenys Bishop

Analytical Services Branch
Methodology Division

Census Analysis Conference 2006
Making the Most of your Census

18–19 July 2006, Canberra, Australia

INQUIRIES

Comments on the research presented in this paper are welcome. However, the contents should not be quoted without the permission of the author(s).

For further information, please contact Dr Glenys Bishop, Analytical Services Branch on Canberra (02) 6252 5140 or email <glenys.bishop@abs.gov.au>.

General and Methodological Issues of the Census Data Enhancement Project

Glenys Bishop
Australian Bureau of Statistics
Locked Bag 10
Belconnen ACT 2616
glenys.bishop@abs.gov.au

Abstract

In August 2005, The Australian Statistician issued a statement of intention outlining plans for the Census Data Enhancement project, the central feature of which is the Statistical Longitudinal Census Data Set (SLCD). This will be based on a 5% population sample and bring together records from the 2006 census with records from the 2011 and subsequent censuses. Names and addresses will not be retained for longer than the census processing period and therefore cannot be used for bringing records from consecutive censuses together. The Statistician also noted that some specified non-ABS data sets would be brought together with the SLCD. In addition, ABS quality studies will be conducted during the census processing period using name and address information.

In this presentation I shall give an overview of this project and describe ABS plans for managing the 5% sample over time, linking records from different censuses, and linking records from the SLCD and other data sets. I shall highlight some of the issues that have arisen in this work, both methodological and of a more general nature. I shall also describe the quality studies that are planned and what assessments of quality we will be making.

Background Information

On 26 April 2005, the ABS put forward a proposal to enhance the use of the Census of Population and Housing (reference 1) . This proposal included the creation of the Statistical Longitudinal Census Dataset (SLCD), and bringing together the SLCD with ABS and specified non-ABS datasets. This discussion paper was put forward for public comment. In conjunction with the proposal, the ABS commissioned a Privacy Impact Assessment (PIA). This PIA had seven recommendations, which the ABS has subsequently addressed.

On 18 August 2005, after consideration of all comments received, the Australian Statistician announced the future of the project, via the Census Data Enhancement - Statement of Intention (reference 2). This project is in line with the legislated function of the ABS to maximise the use, for statistical purposes, of information available to official bodies, by finding further uses for data currently available.

An information paper was released on 8 June 2006, reporting on the status and progress of the project (reference 3). Further information papers are expected to be released in the future.

These papers can all be accessed on the ABS web site, www.abs.gov.au.

Methodology

The classification of "matching" methods is tricky because there is no common terminology. This paper uses terminology believed to be the least confusing, but is not universally applied throughout the literature.

Whatever terminology is used, matching methods belong to two broad groups:

- *Exact matching* is linking records from two different data sets that are believed to belong to the same unit. The result is a dataset of linked units. Exact matching usually involves large data sets, such as administrative data sources or the Census.
- *Statistical matching* involves linking records that do not necessarily belong to the same individual. Rather, the aim is to create a file which accurately represents the population characteristics of both data sets. This is done by modelling the relationship between the variables, using techniques similar to imputation. Statistical matching is usually used when there is little or no overlap between the two datasets (such as linking together two household surveys).

Exact matching methods can be further classified as *deterministic* or *probabilistic*. *Deterministic matching* uses a unique identifier or a set of matching variables, commonly called a matching key. A match is declared when the keys are exactly the same on both data sets. This type of matching can only take place if the data sets have unique, error free identifiers.

In *probabilistic matching* all possible matches or links are evaluated and given a weight based on the likelihood of a match. A match is then declared if the link weight is higher than some predetermined cut-off. This type of matching would occur when there is partial identifying information, but no unique, error free, identifying key. The CDE project is investigating probabilistic exact matching methods.

In November 2005, a paper exploring methodological issues for linking data to form the Statistical Longitudinal Dataset (SLCD) was presented to the Methodological Advisory Committee. It has since been published on the ABS web site as a research paper (reference 4).

In summary, the paper describes the Fellegi-Sunter (reference 5) theory of linking data, and some modifications that more recent authors have considered. In this method, when two records are compared, the values in the same field are compared and a field weight is allocated. The assumption of conditional independence allows field weights to be summed to give a final record-pair weight.

We have since extended the work and software capability to incorporate modifications to simple exact matching of fields. Two that we are working on are value specific field weights and approximate field weights. The former modifies a simple field weight so that more weight is given when a pair of records agree on a relatively infrequent field value. Approximate field weights are used for string texts such as names and other written answers where we assume that two strings match if they are almost the same but not exactly the same.

Quality Studies

A quality study is a study in which a data set is produced by linking the Census with other data sets using names, addresses and other variables. Such data sets can only be formed during the census processing period and will be destroyed at the end of that time. Quality studies will be used to improve ABS products and to investigate linking methodology.

The restrictions on use of these data sets are explained in references 2 and 3.

A major quality study to be performed during the Census processing period will be to assess how well we can link the 5% sample, randomly selected from the 2006 Census, with the 2011 Census. As a model for this, the 2005 Census Dress Rehearsal will be linked to the 2006 Census both with and without names and addresses as matching variables. Linking with names and addresses, while not perfect, will provide a benchmark for assessing linkage quality when statistical techniques are used.

Two quality measures of interest are match rate and match accuracy. First define a match as the pairing of two records, one from each of the files to be linked, which agree on name, address and other variables. When the two files are linked without using names and addresses as linking variables, we would ideally like all the matches to be linked and none of the non-matches. Thus we can define match rate as:

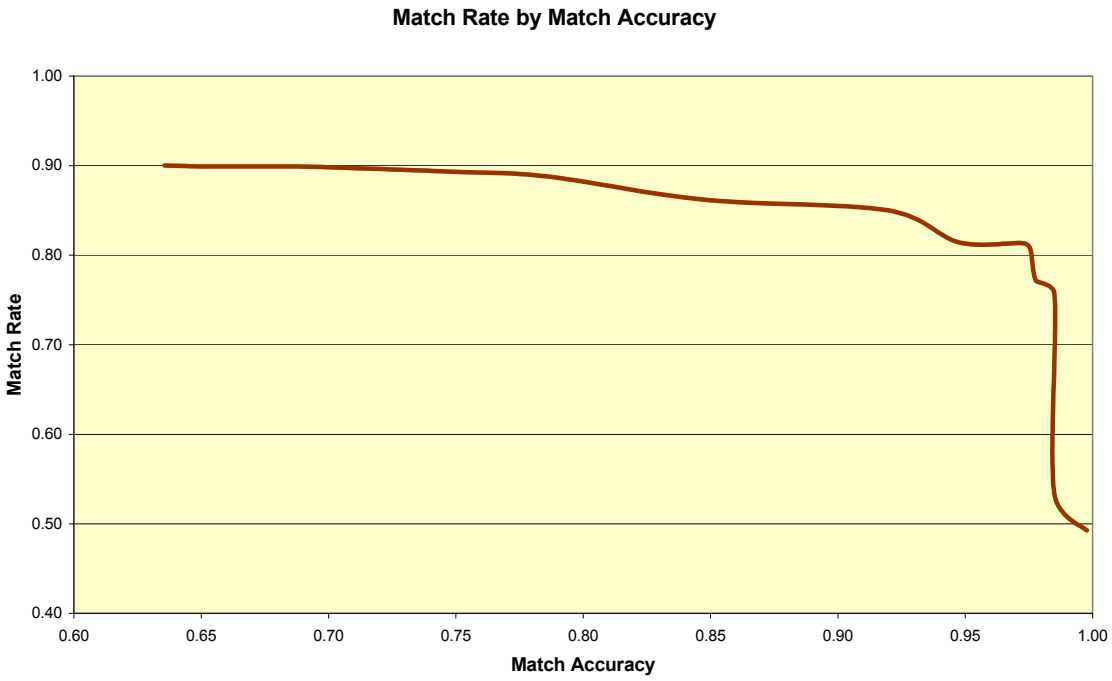
$$\text{match rate} = \frac{\text{number of matches linked}}{\text{number of matches}}$$

and match accuracy as:

$$\text{match accuracy} = \frac{\text{number of matches linked}}{\text{number matches linked} + \text{number non-matches linked}}$$

There is a trade-off between these two measures as can be seen from figure 1.

Figure1. An example of the trade-off between match rate and match accuracy.



Data

Linking techniques are only as good as the data in the files to be linked. Automatic repair mechanisms have been developed for Census questions allowing written responses and these have been adapted for use with names. With the introduction of mesh blocks, methods to code addresses automatically to mesh blocks have been developed. When these automated methods fail, manual repair must be used.

Creating and Maintaining the SLCD

The first wave of the SLCD will consist of a 5% random sample of persons for whom a Census form has been completed in 2006. The second wave will consist of those members of the first wave for whom a Census form has been completed in 2011, augmented by a 5% sample of persons included in the 2011 Census but not included in the 2006 Census. The latter will include children born and immigrants arriving since the 2006 Census. Data collected from the 2011 Census will be combined with the data provided in the 2006 Census. This procedure will continue for future censuses.

The fine details of maintaining this data set have not been completely determined yet.

Confidentiality and Privacy

The ABS is obligated to comply with provisions in the Census and Statistics Act 1905 and the Privacy Act 1988 to respect the privacy of individuals and to protect the confidentiality of their data. The provisions outlined under both these Acts will govern the use and release of data from this project.

Access to non-identifiable unit record data from the SLCD, and the use of the SLCD in conjunction with specified non-ABS data sources, will be subject to the procedures described in reference 3 and will be restricted to access through an ABS data laboratory.

Uses of the SLCD

The potential benefits of the SLCD are substantial. It will provide information on patterns in individual experiences over time and therefore provide insight into the effectiveness of policy or the need for new policy interventions. Examples of studies that could be undertaken are:

- pathways undertaken by migrants in their early years of settlement, particularly in employment;
- links between employment outcomes and education qualifications;
- transitions to higher education and work for young people from low income households; and
- the extent of income and employment mobility.

An important criterion for choosing a method for linking the SLCD from one census to the next will be the effect of the linking method on the analysis of longitudinal data. We are currently engaged in a joint research project with the University of Wollongong to examine models for adjusting for the effects of probabilistic linking. These models will depend to some extent on the types of analyses likely to be performed on the SLCD.

Information is sought from researchers on what they see as potential research questions and likely methods of analysis.

References

1. Australian Bureau of Statistics (2005), Enhancing the Population Census: Developing a Longitudinal View, cat. no. 2060.0, ABS, Canberra

2. Australian Bureau of Statistics (2005), Census Data Enhancement - Statement of Intention, available on the ABS web site, www.abs.gov.au.

3. Australian Bureau of Statistics (2006), Project Update, Statistical Longitudinal Dataset, cat. no. 2062.0, ABS, Canberra

4. Conn, L. and G. Bishop (2006). 'Research Paper: Exploring Methods for Creating a Longitudinal Census Dataset' (Methodology Advisory Committee), cat. no. 1352.0.55.076 , ABS, Canberra

5. Fellegi IP and Alan Sunter (1969), 'A theory for record linkage', Journal of the American Statistical Association, 64(328): 1183 - 1210