Matching in the Australian Census Post-Enumeration Survey

Philip Bell
philip.bell@abs.gov.au

with thanks to Tenniel Guiver (ABS)
What we will cover

- Overview of PES matching
- Computer-assisted matching PES 2006
- Automated data-linking
- Application to PES 2011
Overview of Census operations

- Census collector assigned an area
- Dwellings recorded in collector record book
  - leaves form with household
- Respondents fill out their own form
  - most on paper, some on-line
- Collector picks up form
  - or establishes that on-line form was filled in
Overview of PES operations

- PES provides a probability sample
  - of private dwellings
  - of persons resident in those dwellings
- Matching to Census takes two steps
  - find the dwelling
  - find the person *wherever* they were counted
- Up to 2001 process mainly clerical
- In 2006 computer-assisted
- In 2011 use automated data-linking
  - refer uncertain cases to clerical staff
Information collected by PES

- **Blocklist** of dwellings found in PES
- For each **person** in selected dwellings:
  - **Personal details**
    - name, sex, age, date of birth, marital status, country of birth
  - **Census night address**
    - may not be the current address
  - **Search addresses**: places where the person could have been put on a Census form
Information from Census

- **Collector's Record Book**
  - lists all the dwellings found and whether:
    - occupied (and provided a form)
    - unoccupied
    - non-contact - occupied but no form provided

- **Personal details**
  - for persons in the dwelling on Census night
  - including usual residence address

- Limited personal details for persons usually resident but temporarily absent
Dwelling matching

- Search the collector record book
  - all information stored on computer
  - as images "snippets"
  - confirm address on Census form if available
  - computer may provide multiple CRBs
- If no match, view Census forms
  - check addresses
  - perhaps names of residents
- Can view map
- Can check CRBs of surrounding areas
- Can view PES blocklist
Dwelling match issues

- Two dwellings match to one
  - on PES or on Census
- Search address too unclear
  - e.g. "in Sydney"
  - in which case we impute the probability of a match (from similar search addresses)
Person matching

For each PES person

- Check the dwelling for similar person
- Check persons "temporarily absent"
- Record field match codes
  - eg. for each of name, sex, age, ...
  - eg. "name" has six codes
- Computer assigns person match code
- Computer may ask clerk to check household and family structure
  - easy for clerk, hard for computer
Quality checking

- In 2006, every dwelling *processed twice*
  - two different clerks
- If different results, send to expert clerk
  - to make the final decision
Indigenous communities

- **More difficult** because
  - addresses often incomplete or vague
  - mobile population
  - names change e.g. for religions reasons
- Use **different rules** for matching
  - e.g. less penalty for failing to match on name
- Use **paper listings** of community members in a variety of sort orders
Automatic matching for 2011

- Uses **probabilistic linking**
- Can match PES person at **any** address
  - not just specified search addresses
- Gives list of linked persons from Census
  - either "links" or "possible links"
  - any possible links are **reviewed clerically**
- All confirmed links are treated as matches
- Any person **not linked** after this goes to the **clerical matching** process
What is Probabilistic Data Linking?

File A

File B

Output
Person 1 file A data file B data
Person 2 file A data file B data
...

What is Probabilistic Data Linking?
The Data Linking Process

1. Initial Data
   - Records

2. IT Infrastructure

3. Security Protocols

4. Standardisation
   - Records

5. Blocking and searching
   - Record pairs

6. Comparison
   - Comparison vectors

7. Decision Model
   - Record pairs match status

8. Measurement

The process includes initial data standardisation, blocking and searching, comparison, decision model, and measurement.
Standardisation

- This initial stage ensures that the data on the two files is consistently recorded.
- For example, sex may be recorded as M or F on one file and Male and Female on the other.
Standardisation

- Names
  - name repair
  - nicknames and variants
- Addresses
  - parsing
  - coding (e.g. meshblocks)
- Dates
  - 1 Dunrossil Drive YARRALUMLA ACT 2600
The Data Linking Process

- Initial Data
- Standardisation
  - Blocking and searching
    - Record pairs
    - Comparison vectors
    - Decision Model
      - Record pairs match status
        - Measurement
Blocking

No Blocking

Blocking

e.g. sex, cob, state, CD

2 x 10^12 comparisons

If, say 1000 blocks 80 and 20,000 each

2 x 10^9 comparisons

Blocking variables must be very reliable
as no records outside the block are compared
The Data Linking Process

- Initial Data
- Standardisation
- Blocking and searching
- Comparison
- Decision Model
- Record pairs match status
- Measurement
Variables

2006

Melanie Brown
Female, 16/11/85
1 ABC St, Inner Suburb, Vic 3010
share house 2F, 1M
university student, single
highest schooling Yr 12
highest qual: still studying
no live births
born in Australia
parents born in England
5yrs ago lived:
11 Eve St, Midland, 3091
ancestry English
speaks English

2011

Melanie Searle
Female, 16/11/85
6 DEF Rd, Outer Suburb, Vic 3215
couple 1M, 1F
pharmacist, married
highest schooling Yr 12
highest qual: bachelor degree
no live births
born in Australia
parents born in England
5yrs ago lived:
1 ABC St, Inner Suburb, Vic 3010
ancestry English
speaks English
### Improving the Distinguishing Power of Names

<table>
<thead>
<tr>
<th>Name</th>
<th>Sex</th>
<th>Common Male Name</th>
<th>Uncommon Male Name</th>
<th>Common Female Name</th>
<th>Uncommon Female Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jonathon</td>
<td>M</td>
<td>John</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Tenniel</td>
<td>M</td>
<td>------</td>
<td>Tenniel</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Jenny</td>
<td>F</td>
<td>------</td>
<td>------</td>
<td>Jennifer</td>
<td>------</td>
</tr>
<tr>
<td>Tenille</td>
<td>F</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>Tenille</td>
</tr>
</tbody>
</table>
Other Variables

- **Surnames**
  - Common/Uncommon

- **Country of Birth**
  - set to missing if born in Australian
  - otherwise use 2 digit region of birth

- **Indigenous**
  - more likely than an unmatched pair agrees on being non-indigenous
Comparisons

- When comparing a record from file A with one from file B, we look at the same fields on each.
  
<table>
<thead>
<tr>
<th></th>
<th>dob</th>
<th>MB</th>
<th>cob</th>
</tr>
</thead>
<tbody>
<tr>
<td>record a</td>
<td>16/3/79</td>
<td>1031812</td>
<td>Australia</td>
</tr>
<tr>
<td>record b</td>
<td>16/3/79</td>
<td>1390101</td>
<td>Australia</td>
</tr>
<tr>
<td>comparison</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- Combine the results of field comparisons to create the final, record pair comparison value.
  - The bigger this is the stronger the chance of a match.
- Underlying mathematical theory in this step.
Field Comparison Functions

- SHACKELFORD  SHACKELFOBD (0.96)
- SHACKLEFORD  SMACKLEFORD (0.95)
- ITMAN          SMITH (0.57)

Similarly, we can use partial agreement on numeric fields
Probabilities and Weights

- **Inputs**
  - probability that the fields agree given that the records belong to the same person
  - probability that the fields agree given that the records belong to different people

<table>
<thead>
<tr>
<th></th>
<th>same person</th>
<th>different</th>
<th>agree weight</th>
<th>disagree weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>0.999</td>
<td>0.5</td>
<td>0.3</td>
<td>-2.7</td>
</tr>
<tr>
<td>date of birth</td>
<td>0.95</td>
<td>0.0015</td>
<td>2.8</td>
<td>-1.3</td>
</tr>
</tbody>
</table>

- Different variables have different distinguishing power
What Have We Got So Far?

- Set of matched records with an associated comparison weight
- Lots of these, high weights, low weights and in-between
- Many to many comparisons
The Data Linking Process

- Initial Data
- Standardisation
  - Blocking and searching
- Comparison
  - Comparison vectors
- Decision Model
  - Record pairs match status
- Measurement

Records
Record pairs
Results of Comparisons

Frequency

Comparison Weight

Non-Links

Possible links

Matches

Links

Incorrect non-links

False links

Non-matches
Clerical review these record pairs.

Upper clerical threshold

Set of all record pairs

Clerical review these record pairs

Automatically accept all these record pairs as links

Automatically reject all these record pairs as non links
The Home Stretch

Initial Data → Records → Standardisation → Records → Blocking and searching → Record pairs → Comparison → Comparison vectors → Decision Model → Record pairs match status → Measurement
## Characterising the Results

<table>
<thead>
<tr>
<th></th>
<th>Matches</th>
<th>Non-matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links</td>
<td><img src="image1" alt="Matched Links" /></td>
<td><img src="image2" alt="Non-matched Links" /></td>
</tr>
<tr>
<td>Non-links</td>
<td><img src="image3" alt="Matched Non-links" /></td>
<td><img src="image4" alt="Non-matched Non-links" /></td>
</tr>
</tbody>
</table>
PES Quality Study

- A study to investigate the feasibility of using data linking to automate the clerical matching process
- Found to introduce efficiencies and find extra matches
Benefits of automated match

- Works with vague search addresses
- Finds additional matches
  - when tested on PES 2006
- Better use of clerical resources
  - used on the difficult cases
  - and to confirm the questionable cases
- Can ensure high match accuracy
  - only link when match is clear
    - e.g. aim for ~60,000 automated links
    - *clerically review* the remaining ~30,000
      (to find ~20,000 extra matches).
The Future: PES 2011

- Aim to integrate automated data linking into the PES clerical process
- **Batch** operation
  - one state at a time as processing completed
  - supplement with a final nationwide search.
- **System design** underway
Linking software: FEBRL

- ABS used free software FEBRL
  - author Peter Christen (ANU)
  - does blocking, linking etc.
  - written in PERL
  - slow

- Huge datasets
  - 80,000 PES persons, 20,000,000 Census
  - speed depends on platform
  - need lots of RAM memory
ABS changes to FEBRL

- New comparators
  - e.g. how close are two ages
    - 1 year age difference is a good match
  - special treatment of "missing"

- Speed up indexing
  - so can quickly access data for a record

- Add clerical review
  - displays relevant data for two records
  - allows clerk to enter "match" or "no match"

- All done with older version of FEBRL