An application of probabilistic matching

Abowd and Vilhuber (2004), JBES

An example

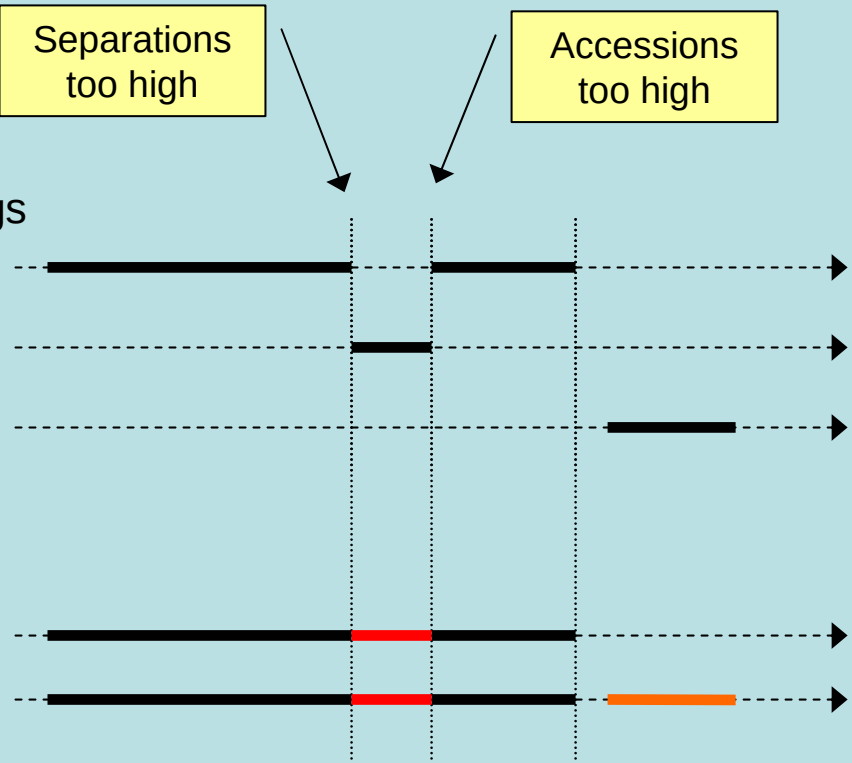
Abowd and Vilhuber (2004), JBES: "*The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers*"

- Approx. 500 million records (quarterly wage records for 1991-1999, California)
- 28 million SSNs

SSN Name editing

Example

Coded Name	Coded SSN	EIN	Earnings
Leslie Kay	1	A	$10
Leslie Kay	2 1	A	$10
Lesly Kai	3 1	B	$11



1's tenure with A:

1's employment history

A&V: standardizing

- Knowledge of structure of the file:
  -> No standardizing
- Matching will be within records close in time -> assumed to be similar, no need for standardization
- BUT: possible false positives -> chose to do an weighted unduplication step (UNDUP) to eliminate wrongly associated SSNs

A&V: UNDUP

SSN	*UID*	*First*	*Middle*	*Last*	*Earn*	*YQ*
123-45-6789	58	John	C	Doe	25678	93Q1
123-45-6789	58	John	C	Doe	26845	93Q2
123-45-6789	59	Jon	C	Doe	24837	94Q4
123-45-6A89	60	Robert	E	Lee	7439	93Q1

A UID is a unique combination of SSN-First-Middle-Last

A&V: UNDUP (2)

SSN	*UID*	*First*	*Middle*	*Last*	*Earn*	*YQ*
123-45-6789	58	John	C	Doe	25678	93Q1
123-45-6789	58	John	C	Doe	26845	93Q2
123-45-6789	59	Jon	C	Doe	24837	94Q4
123-45-6789	60	Robert	E	Lee	7439	93Q4
123-45-6789	60	Robert	E	Lee	7439	94Q1

Conservative strategy: Err on the side of caution

Matching

- Define match blocks
- Define matching parameters: marginal probabilites
- Define upper T_u and lower T_l cutoff values

Record Blocking

- Computationally inefficient to compare all possible record pairs
- Solution: Bring together only record pairs that are LIKELY to match, based on chosen blocking criterion
- Analogy: SAS merge by-variables

Blocking example

- Without blocking: **AxB** is 1000x1000=1,000,000 pairs
- With blocking, f.i. on 3-digit ZIP code or first character of last name. Suppose 100 blocks of 10 characters each. Then only 100x(10x10)=10,000 pairs need to be compared.

A&V: Variables and Matching

- File only contains Name, SSN, Earnings, Employer
- Construct frequency of use of name, work history, earnings deciles
- Stage 1: use name and frequency
- Stage 2: use name, earnings decile, work history with employer

A&V: Blocking and stages

- Two stages were chosen:
  - UNDUP stage (preparation)
  - MATCH stage (actual matching)
- Each stage has own
  - Blocking
  - Match variables
  - Parameters

A&V: UNDUP blocking

- No comparisons are ever going to be made outside of the SSN
- Information about frequency of names may be useful
- Large amount of records: 57 million UIDs associated with 28 million SSNs, but many SSNs have a unique UID

$\Rightarrow$Blocking on SSN

$\Rightarrow$Separation of files by last two digits of SSN (efficiency)

A&V: MATCH blocking

- Idea is to fit 1-quarter records into work histories with a 1-quarter interruption at same employer

$\Rightarrow$Block on Employer – Quarter

$\Rightarrow$Possibly block on Earnings deciles

A&V: MATCH block setup

```
# Pass 1:
BLOCK1 CHAR SEIN SEIN
BLOCK1 CHAR QUARTER QUARTER
BLOCK1 CHAR WAGEQANT WAGEQANT
# follow 3 other BLOCK passes with identical setup
#
# Pass 2: relax the restriction on WAGEQANT
BLOCK5 CHAR SEIN SEIN
BLOCK5 CHAR QUARTER QUARTER
# follow 3 other BLOCK passes with identical setup
```

Determination of match variables

- Must contain relevant information
- Must be informative (distinguishing power!)
- May not be on original file, but can be constructed (frequency, history information)

A&V: UNDUP match variables

```
# Pass1
MATCH1 NAME_UNCERT namef 0.9 0.001 700
MATCH1 NAME_UNCERT namel 0.9 0.02 700
MATCH1 NAME_UNCERT namem 0.9 0.02 700
MATCH1 NAME_UNCERT concat 0.9 0.02 700
# Pass 2
MATCH2 ARRAY NAME_UNCERT fm_name 0.9 -.02 750
MATCH2 NAME_UNCERT namel 0.9 0.001 700
MATCH2 NAME_UNCERT concat 0.9 0.02 700
# and so on…
```

A&V: MATCH match variables

```
# Pass1
MATCH1 CNT_DIFF SSN SSN 0.9 0.000001 5
MATCH1 NAME_UNCERT namef namef 0.9 0.02 700
MATCH1 NAME_UNCERT namel namem 0.9 0.02 700
MATCH1 NAME_UNCERT namel namel 0.9 0.001 700
# Pass 2
MATCH2 CNT_DIFF SSN SSN 0.9 0.000001 5
MATCH2 NAME_UNCERT concat concat 0.9 0.02 700
# Pass 3
MATCH3 UNCERT SSN SSN 0.9 0.000001 700
MATCH3 NAME_UNCERT namef namef 0.9 0.02 700
MATCH3 NAME_UNCERT namem namem 0.9 0.02 700
MATCH3 NAME_UNCERT namel namel 0.9 0.001 700
```

and so on…

Adjusting P(agree|M) for relative frequency

- Further adjustment can be made by adjusting for relative frequency (idea goes back to Newcombe (1959) and F&S (1969))
  - Agreement of last name by Smith counts for less than agreement by Vilhuber
- Default option for some software packages
- Requires strong assumption about independence between agreement on specific value states on one field and agreement on other fields.

A&V: Frequency adjustment

- UNDUP:
  - none specified
- MATCH:
  - allow for name info,
  - disallow for wage quantiles, SSN

Marginal probabilities: better estimates of P(agree|U)

- P(agree | U) can be improved by computing random agreement weights between files α(A) and β(B) (i.e. **AxB**)
  - # pairs agreeing randomly by variable X divided by total number of pairs

Error rate estimation methods

- Sampling and clerical review
  - Within L: random sample with follow-up
  - Within C: since manually processed, “truth” is always known
  - Within N: Draw random sample with follow-up. Problem: sparse occurrence of true matches
- Belin-Rubin (1995) method for false match rates
  - Model the shape of the matching weight distributions (empirical density of R) if sufficiently separated
- Capture-recapture with different blocking for false non-match rates

Analyst Review

- Matcher outputs file of matched pairs in decreasing weight order
- Examine list to determine cutoff weights and non-matches.

A&V: Finding cutoff values

- UNDUP:
  - CUTOFF1 7.5 7.5
  - CUTOFF2 8 8
  - Etc.
- MATCH:
  - CUTOFF1 18 18
  - CUTOFF2 12 12
  - CUTOFF 10 10
  - Etc.

A&V: Simulated matcher output

RESULT	RECNUM	WGT	SSN	NAMEF	NAMEM	NAMEL
[UA]	504	-999.99	382661272	WILL		TARY
[UB]	2827	-999.99	384883394	RICHARD		PHOUK
[UB]	392	-999.99	335707385	MONA		LISA

RESULT	RECNUM	WGT	SSN	NAMEF	NAMEM	NAMEL
[CA]	351	3.66	333343734	DONNA	L	DUK
[CB]	1551	3.66	333383832	MARGEN	L	PRODUCT

RESULT	RECNUM	WGT	SSN	NAMEF	NAMEM	NAMEL
[MA]	43	32.76	444444441	LUKE		UPP
[MB]	169	32.76	444444447	LUKE		UPP

Post-processing

- Once matching software has identified matches, further processing may be needed:
  - Clean up
  - Carrying forward matching information
  - Reports on match rates

Generic workflow (2)

- Start with initial set of parameter values
- Run matching programs
- Review moderate sample of match results
- Modify parameter values (typically only m_k) via ad hoc means