

An Empirical Comparison of Two Record Linkage Procedures

S. Gomatam, R. Carter, M. Ariet,
G. Mitchell

Outline

- Ideas of Exact/Direct matching
- Probabilistic matching, AUTOMATCH
- Deterministic strategies, SDS
- Performance
- Comparison
- Discussion

General Idea

- Two files \rightarrow A and B with n_a and n_b records drawn from the same population.
- Each record has values on various fields (or identifiers). Files share common fields but no unique identifier.
- $n_a \times n_b$ record pairs in $A \times B$ have to be identified as matches (M) or non-matches (U).
- k common fields, $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_k)$ is agreement vector.
- Classification strategies either probabilistic or deterministic.

Probabilistic Strategy

Fellegi and Sunter (1969, JASA)

- M: set of matches, U: set of non-matches, γ : agreement vector
- $m(\gamma) = P(\gamma | (a,b) \text{ in } M)$, $u(\gamma) = P(\gamma | (a,b) \text{ in } U)$. m/u is large if pair in M, small if pair in U.
 $w = \log(m/u)$
- Decisions on match/non-match based on composite weight w
- Under conditional independence,
 $w = \sum \log(m_i/u_i) = \sum w_i$

- Can classify pair as either match (A1), uncertain (A2), non-match (A3)
- “Optimal strategy”: Fix μ =false match rate, λ =false non-match rate, minimize classification in A2
- Method: Order γ using w . Decide cutoff weight thresholds to achieve μ , λ . Large in A1, small in A3, rest in A2
- Estimate weights using knowledge of error rates from prior data, or ...

Blocking

- Method to reduce number of comparisons actually carried out.
- Only record pairs in subsets of files that agree on “blocking” variables are examined.
- Pairs not in same block are considered non-matches, so possible inflation of false match rate. Minimize by using multiple passes with different blocking variables.

AUTOMATCH

Jaro (1989, JASA; 1995 Stat Med)

- Implementation of prob. Inkg
- Weights estimated via EM algorithm
- Uses linear programming to decide unique matches
- Allows partial agreements
- Requires: specification of blocking and matching variables, initial values for the m and u prob.s, cut-offs on weights

Deterministic Strategy

- Consider “all or nothing” matches on identifiers. Total agreement indicates match, else non-match.
- Implementations ad hoc
- Consider systematic strategy, classify based on the most reliable subset, consider less reliable collections at subsequent stages. For eg., start with (lastname, firstname, date of birth), next consider (lastname, date of birth).

Implementation of stepwise deterministic strategy

- Stepwise matching on subsets
- Unique matches retained, residuals to next stage
- Continue till “comfort level” isn’t exceeded
- Implementation via SAS macro

Why Compare?

Probabilistic

- Formal method
- Invest to understand and implement
- Need info. and method to come up with good weights
- Assume independence

Deterministic

- Informal, ad hoc
- Intuitive, easier to implement
- May be able to use intuition to decide "decent" sequence
- Assume experience practitioner

How compared

- DOEd database with info. on special education placements and socio-demographic variables for school-going children
- RPICC database has medical and some socio-demographic info. on infants
- Compare strategies on 1156 RPICC records that have unique SSN matches in DOEd. Link these with 7437 DOEd records (1% added). [Also did 0.5% (4296) and 5%-added (32549).]
- Identifiers available to both strategies: last name, first name, middle name, NYSIIS codes of names, date of birth, race, county number, and sex.

On NYSIIS codes

Phonetic coding system to address problems of alternative spellings of names. Rules applied to names to create codes (see Newcombe, 1988).

Tradeoff between bringing together alternative forms of the same name, and achieving high level discriminating power (NYSIIS vs. Soundex)

AUTOMATCH results

- Ran 5 passes. Matching variables used some subset of: NYSIIS codes of first, middle and last names, last name, date of birth (with variation of 7 days allowed in pass 2)
- Blocking variables used for first 2 passes {NYSIIS code of last name}, for 3rd and 4th passes {Date of birth}, and for final pass {Race, County number, Sex}.

Breakup of matches in 5 AUTOMATCH passes

Pass number	1	2	3	4	5	Total
0.5%-added True	657	3	181	8	194	1043
Total	664	3	182	8	207	1064
1%-added True	657	4	184	8	189	1042
Total	664	4	186	9	210	1073
5%-added True	658	3	181	8	194	943
Total	686	5	159	9	183	1042

Breakup of pairs under AUTOMATCH linkage (1%-added)

Decision\Truth	Match	Non-match	Total
Match	1042	31	1073
Non-match	113	8,595,985	8,596,099
Total	1156	8,596,016	8,597,172

SDS results

- Ran 4 segments of SDS linkage:
I=matches with county number included;
II= matches w/o county number;
III=matches for last name changes;
IV=matches for first/middle name problems

Breakup of matches in 4 segments

Segment number (No. of var.s)	I (7)	II (6)	III (6)	IV (5)	Total
0.5%-added True	326	57	169	215	767
Total	326	57	169	216	768
1%-added True	326	57	168	214	765
Total	326	57	168	215	766
5%-added True	323	56	168	207	754
Total	324	56	168	210	758

Breakup of pairs under SDS linkage (1%-added)

Decision\Truth	Match	Non-match	Total
Match	765	1	766
Non-match	391	8,596,015	8,596,406
Total	1156	8,596,016	8,597,172

Accuracy statistics

Sample	Statistic	AUTOMATCH	SDS
0.5%-added	Match rate	0.9204	0.6644
	PPP	0.9803	0.9987
	Sensitivity	0.9022	0.6635
	Specificity	0.9999	0.9999
1%-added	Match rate	0.9282	0.6626
	PPP	0.9711	0.9987
	Sensitivity	0.9014	0.6618
	Specificity	0.9999	0.9999
5%-added	Match rate	0.9014	0.6557
	PPP	0.9050	0.9947
	Sensitivity	0.8157	0.6522
	Specificity	0.9999	0.9999

Comparison

- Match rates: AUTOMATCH -> 93%, SDS -> 66% (92-66; 90-66)
- Specificities practically the same. AUTOMATCH has much higher sensitivity than SDS (90 vs. 66)
- Positive predictive probability for AUTOMATCH is much lower (92.8 vs. 99.6)
- As sample size increases: sensitivity and PPP of AUTOMATCH linkages drops by about 0.08; PPP for SDS is larger by about 0.02-0.09 over all 3 samples; SDS practically unaffected by sample size, AUTOMATCH performance deteriorates

Discussion

- Subjectivity in both methods – AUTOMATCH: choice of blocking var.s, matching var.s, cut-offs, use of clerical r/w; SDS: choice of matching var.s and sequence, decision on when to stop.
- AUTOMATCH does better when greater sensitivity or overall accuracy is desired (A2 and clerical review add extra value).
- SDS better for situations where high PPP is important.
- Additional features AUTOMATCH: ease of clerical review, info. theoretic char. cmp., leeway in dates, value-specific prob.s for m and u; SDS: informal inc. of investigator's knowledge.
- Hybrid approaches used in practice.
- Designed experiment/simulation to study comparison in detail?

Caveat: One implementation of AUTOMATCH and SDS, single data set.

Reference

- This work available in paper “An Empirical comparison of record linkage procedures,” (2002), *Statistics in Medicine*, 21, 1485-1496
- For PDF or postscript version write to sgomatam@niss.org