

Practical and Theoretical Considerations for Linking Survey Data with Other Sources

Michael D. Larsen

Iowa State University
Department of Statistics
Center for Survey Statistics & Methodology
larsen@iastate.edu

Friday, February 29, 2008

Conference on Linking NSF SED/SDR Data to Scientific Productivity Data

Thanks to organizers

- ▶ Donna Gither
- ▶ Jinyoung Kim
- ▶ Gerald Marschke

for organizing and for the invitation to present ideas.

Research project overview

- ▶ Women in science are underrepresented, earn less, and are less likely tenured than men. Why? Very important research question.
- ▶ In addition to SDR information, match to publications and patents to get better information.
- ▶ Great idea – the data could be very predictive of differences within fields of study.

Research project comments

- ▶ Note: SDR has interesting data in restricted use data set on job responsibilities, PhD training, etc.
- ▶ Note: some journals not indexed (*Survey Methodology*, *JOS*)
- ▶ Note: any rating system for journals has its limitations/biases.
- ▶ What about match to funding sources for some fields?

Research project comments, 2

- ▶ Other issues: family – do unequal roles and expectations of men and women in families explain observed differences? E.g., women more likely to have a spouse in academics and take more time initially for children.
- ▶ Other issues: aggressiveness – men more likely to try to relocate, get bigger retention offers.
- ▶ Also important: service, editorial positions, funding, awards, students, teaching, etc. – difficult to quantify/get data.
- ▶ In general, concluding causation is very hard.

Overview of record linkage operations

1. Determine matchable variables in files.
2. Examine limitations of the variables: level of quality/errors, missingness, uniqueness, and stability over time.
3. Define how to eliminate totally unlikely matches.
4. Do comparisons on remaining pairs.
5. Create composite evaluation of match.
6. Make decisions.
7. Examine results. Refine procedure.

1. Matchable variables for this application

- ▶ First, middle, last name. See comments by Winkler – some names will be much more informative than others.
- ▶ ZIP code, State of residence/employer– are these the same in both files? – Perhaps consider metro regions. – Some locations will be much more informative than others.
- ▶ Age? – Can this be inferred or created from DOB? It is in SDR.
- ▶ Department versus field – Is there anything to strong agreement here?
- ▶ Title: assistant, associate, full professor?
- ▶ Coauthors, coinventors, etc. – This could be useful, especially as a positive indication of match – 'social network'

2. Limitations of variables

- ▶ Level of quality/errors: do you have any information on errors in SDR? in patent citations? in publication index?
- ▶ Missingness – middle name or initial; others?
- ▶ Uniqueness – several Asian names will not be unique; some ZIPs and States will be quite common. Field and institution could be more unique.
- ▶ Stability over time – academic publications can take a long time to appear – are affiliations stable? Look \pm a couple of years.

3. Eliminate totally unlikely matches: Blocking

- ▶ A 'block' is defined as a set of records with something in common – e.g., first letter last name, same state/region, same institution.
- ▶ Assume records in different blocks are not matches – reduces computation.
- ▶ Works pretty well in census applications.
- ▶ Sometimes linkage results can be sensitive to blocking criteria – try different sets.
- ▶ Could be used in more flexible ways – if disagree on 3 of the following 6, then not a possible match.

4. Do comparisons on remaining pairs.

- ▶ Exact agreement versus partial agreement.
- ▶ Location: what year? how exact (metro area)?
- ▶ Names: see Winkler for comparison metrics.
- ▶ Define rules for missing entries.

5. Create composite evaluation

- ▶ This often is deterministic – if such and such criteria are satisfied, then call it a match.
- ▶ Fine if few errors due to errors in matching variables.
- ▶ Fine if legitimate changes can be addressed (affiliation for a publication might not match current affiliation; employer and residence states can be different).

Other methods for composite

- ▶ Other procedures give weights ($+a_x$ for agree, $-d_x$ for disagree on variable x) that are added together.
- ▶ Agreement on some names/locations is more important than others
- ▶ Disagreement on some variables is more important than others
- ▶ You can mix required fields (e.g., blocking) with scoring – scores within blocks.
- ▶ Census applications: Latent class models can be used.

Latent class methods: Fellegi and Sunter 1969

- ▶ $P(M|x) = P(M)P(X|M)/P(X)$ – an application of Bayes' theorem
- ▶ Latent class model: $P(X) = P(M)P(X|M) + P(U)P(X|U)$
- ▶ $P(X|M)$ and $P(X|U)$ simplify under latent class assumption
- ▶ All components of RHS can be estimated via MLE
- ▶ Determines weights via the data themselves.
- ▶ See also Winkler's work and my papers.
- ▶ Will it work here?

6. Make decisions

- ▶ Sometimes deterministic for either declare link or for declare non link.
- ▶ Blocking can declare non links.
- ▶ Need cut off values for composite weights – can be based on empirical evaluation.
- ▶ Latent class approach provides estimates of probabilities of match and probabilities of errors.

7. Examine results

- ▶ Sample of records and clerical evaluation
- ▶ How many records? More is better. If $0.05 = 5\%$ error rate and $n = 150$, MOE is about $0.036 = 3.6\%$. If $0.10 = 10\%$ error rate and $n = 150$, MOE is about $0.048 = 4.8\%$.
- ▶ Which records? Those that are hard to match but are likely matches.
- ▶ **Suggestion:** Stratify the population – take some from different groups. In an academic setting, stratify by field and, if possible, race/ethnicity. Sample less in easy to match cases – but sample some there. Sampling theory then produces estimates for population of declared matches.

7. Refine procedures

- ▶ Fewer missed matches – use less stringent blocking; less stringent agreement criteria.
- ▶ Fewer false positives – use more stringent blocking; more stringent agreement.
- ▶ Change blocking; change comparison rules.
- ▶ Try different approaches and compare/contrast.
- ▶ **Suggest:** think about options like a designed experiment – look also at interactions between choices.

Impact on analysis

- ▶ If keep best matches, then clearest analysis, but there could be bias. Asian names in big cities and women will be harder to match.
- ▶ If allow in a few less likely matches, then possibly more errors in matching; potentially extreme measurement error.
- ▶ **Suggest** keeping measurement(s) of quality along with match status. Run analysis at different levels of matching. Report sensitivity to matching assumptions.
- ▶ **Suggest** stratified random sample of cases – this allows clearer inference.
- ▶ Statistical procedures could be used to try to identify errors – see Scheuren and Winkler (and Lahiri and Larsen) on regression of linked files.

Multiple studies possible

- ▶ One could publish and report on various matching studies
- ▶ and analyses of matched data.
- ▶ Large clerical review always will be desired – but it is expensive and time consuming.
- ▶ Suggest stratified random sample of cases for review.

Thoughts on confidentiality

- ▶ Does linking the SDR to other data sources produce a confidentiality concern?
- ▶ Possibly need to limit access to linked data – only certain users should have linked information.
- ▶ Or need to take steps to preserve confidentiality – details of published papers, name of institution and/or dept are very specific to individual – only certain variables should be released.
- ▶ Idea for top coding some variables might be reasonable solution. But interactions of variables could still be quite informative.
- ▶ A confidentiality evaluation would be a substantial project on its own.

Thanks

- ▶ And good luck!
- ▶ larsen@iastate.edu