Discussion of “Would the Real Steve Fienberg Please Stand Up: Getting to Know a Population from Multiple Incomplete Files” Invited Papers, Social Statistics Section

Michael D. Larsen[1]
[1]Department of Statistics, George Washington University
6110 Executive Blvd, Suite 750, Rockville, MD 20852
mlarsen@bsc.gwu.edu

Abstract

This proceedings paper presents discussion of four papers presented in the invited paper session, Would the Real Steve Fienberg Please Stand Up: Getting to Know a Population from Multiple Incomplete Files, which was part of the Social Statistics Section program (Session 572) of the 2013 Joint Statistical Meetings in Quebec, Montreal, Canada.

Key Words: Bayesian methods, Blocking, Clustering, Conditional classifiers, Coverage measurement**,** De-duplication, False match rates, False nonmatch rates, Post stratification, Record linkage, Smoothing.

1. Introduction

This proceedings paper presents discussion of four papers presented in the invited paper session, Would the Real Steve Fienberg Please Stand Up: Getting to Know a Population from Multiple Incomplete Files, which was part of the Social Statistics Section program (Session 572) of the 2013 Joint Statistical Meetings in Quebec, Montreal, Canada.

The four papers present innovative methods and apply to interesting data sets. Brief comments are included here. It is beyond the scope of this discussion to comment thoroughly on algorithms, applications, and results.

There are some common themes present in these four talks:

- Multiple incomplete files, or one file with duplicates
- Integration/de-duplication is needed to get a better picture/analysis
- Statistical models/methods are needed to bridge the gap between what is known and what the analyst wants to know

These papers address important topics with many applications:

- A vast array of topics fall into categories (e.g., health and medicine, economics and labor force, demographics) illustrated in these talks.
- The number of large electronic databases available for use in analyses such as these is large and will continue to grow very quickly.
- As costs of original data collection continue to escalate and face challenges, methods that can use data from multiple sources toward an end will increase in value.

Some words of caution apply to these four talks and in general to applications in this area:

- Input data quality matters a lot and all files are not created equal
- Example: U.S. Census Bureau has been working (for decades, on/off again) to use administrative records to count the population and supplement the Decennial Census. Pre-processing files to clean up names, addresses, dates, etc., is critical for any successful application. Commercial files have severe limitations, including a tendency to buy data from each other.
- Validation of results is challenging, but very important. Sensitivity analysis to departures from assumptions should accompany efforts at combining data sets and analyzing the resulting files.

The next four sections discuss the papers presented in this JSM invited papers session.

2. Sam Ventura, Deduplicating text records using clustering and aggregation of conditional classifiers

The presentation by Sam Ventura (Ventura and Nugent 2003) concerns one large file on inventors and their patents. Similar applications could arise looking at authors of academic papers. Deduplication also would arise in Census records. The problem in a nut shell is to link patents to inventors. This is a many-to-one match: each inventor has many patents. The authors have training data that consists of data with labeled links between patents and inventors.

Several types of variables are available for matching in the patent database. The paper describes similarity measures for the variables. Options are given for different data types (long strings, short strings, numbers, or lists of names). Training data are used in the process to train classifiers. The procedure aggregates multiple classifiers and is referred to as “forests of random forests”. The output of the procedure is a distance measure that is used in agglomerative clustering. Clustering is done separately for cases with and without middle names, because the existence of a middle name impacts the ability to discriminate among inventors in the patent database. As a whole the proposed scheme is an innovative approach to a number of details that reflects careful thought about data.

In order to address the large file size, blocking is used. That is, records are eligible for matching only if the inventors on two patents have the same first letter of the last name. A second detail that facilitates handling the large file size is ‘comparison storage’ which creates comparison results between each pair of unique entries and then uses results as a look up table. This reduces comparisons, including string comparisons, while processing the entire file. A third modification that addresses large file size is the forests of random forests approach.

Here are some ideas that could be studied in reference to the methods proposed in this paper.

- Are errors introduced by blocking? How does the procedure handle last name changes from, for example, a maiden name to a married name? Could a block be based on “first letter of last name or some other important agreement” in order to avoid errors due to name changes?

- How much training data really are needed? Training data are expensive as they often require human clerical review. Are training data ‘transferable’ to another application (another patent area)?
- Can you eliminate many pairs a priori as simply not reasonable? Do any pairs match exactly and can they be a prior linked? That is, can you be sure of some decisions without having to put the record pairs through the classifier? Doing so could save time without impacting accuracy to any great amount.
- How do Forests of Random Forests perform in comparison to Bayesian Classification trees?
- Do agglomerative and divisive clustering procedures produce similar results? Agglomerative clustering begins with all observations in singleton clusters and then joins them together. Divisive clustering begins with all observations in one super cluster and then divides into subsets.
- Do robust distance measures produce substantially different results for any subset of record pairs than the chosen distance measures?

3. Rebecca Steorts, Getting to know the population from multiple incomplete files

The presentation by Rebecca Steorts (Hall, Steorts, and Fienberg 2013) concerns linking many files together simultaneously. The example uses six (6) waves of the National Long Term Care Survey, which has approximately 20,000 respondents per wave. Simply taking procedures for matching two files together and applying it to each pair of files can be done, but it runs into a problem of possible non-transitivity: even if *a* and *b* match in comparison of files A and B and *b* and *c* match in comparison of files B and C, there is no guarantee that a match of files A and C will put *a* and *c* together.

The authors propose an innovative solution to the non-transitivity problem. The population covering all waves of the survey is deemed the Latent population. The Latent population contains all unique individuals from the composite of surveys with true values of all measured variables. The files themselves then consist of observations on subsets with possible duplicates and errors in variables. Records from the several files then are linked to the latent population.

The problem is given a Bayesian formulation. The Bayesian formulation allows simulation of both model parameters and links to the Latent population. A hybrid MCMC algorithm (Metropolis-Hastings for jumps in linkages) is given for computing. The Bayesian approach enables flexible summaries of results. Indeed, the proposed method can handle record linkage, de-duplication, and capture-recapture size estimation all at once. Conceptually, these problems are similar, so it is not surprising that a proper formulation of the overall problem can lead to a solution to all of them.

There are a few questions that could receive further investigation.

- Blocking: blocking is used to reduce computation – how sensitive are results to blocks? Can you compare results with different blocking criteria?
- Can you eliminate many pairs a priori as simply not reasonable? Do any pairs match exactly and can they be a prior linked? That is, can you be sure of some decisions without having to put the record pairs through the algorithm? Doing so

could save time without impacting accuracy to any great amount. It also could have the advantage of helping algorithm performance. In mixture or latent class models, when you can use some labeled training data, it can effectively ‘identify’ the classes, thereby helping with interpretation and monitoring algorithm performance.

- Additional scenarios can be considered to approximate challenges that might be seen in some applications. Other scenarios to consider could include: a lot of noise records with no links; more errors in matching variables; files with different data quality; missing data; and files of (vastly) different size.

4. Tom Mule, Bayesian record linkage models for Census coverage measurement matching

Tom Mule (Mule and Imel 2013) present work on the Fellegi-Sunter algorithm done in a Bayesian formulation. The work extends previous work by Larsen (2009). Their application is estimation in the 2010 Census Coverage Measurement study that was used to evaluate the accuracy of the 2010 decennial census. Prior distributions are added to the statistical likelihood function in order to ‘go Bayesian’. Computing uses Gibbs sampling with the links between record pairs considered to be missing data.

In order to deal with the size of the files and comparisons, blocking is used to reduce the total number of comparisons. The decennial census counts the population by address. As a result, blocking is by household and blocks are quite small. Sometimes there is only one person in a block in the Census and in the post enumeration survey that is linked to the Census.

Names are compared using string comparator metrics. The work in this paper utilizes an innovation by categorizing the results of string comparison into four levels based on the degree of similarity plus a fifth category for missing information. Other fields, such as age, allow agreement, disagreement, and missing. Another person who has investigated alternative use of string comparisons in record linkage is William Winkler, also at the U.S. Census Bureau.

In the statistical models they use, the authors can specify that the latent class parameters are the same across blocks, or they can allow them to vary. One-to one (1-1) matching is enforced post hoc (linear sum assignment procedure). Larsen (2009) suggested ideas for incorporation 1-1 matching directly into the likelihood function and sample space. Larsen (2009) also suggested algorithms for computing in this context.

The authors use the current implementation of the algorithm to identify cells of the patterns of agreement table that have large deviations in the predicted probability of matching from the actual rate. They are able to do this calculation because they have labeled training data available.

As mentioned previously, this brief discussion does not address performance of methods or application to data in any detail. The interested reader is referred to the proceedings papers and other publications by the relevant authors.

There are a few questions that could receive further investigation

- Is there any way to loosen blocking to see what happens? Instead of household, what if you use street or cluster of house numbers on a street? A looser blocking criterion increases the amount of computation and could introduce false matches. On the other hand, the strict blocking currently being used could produce false non matches (missed matches).
- Can you eliminate many pairs a priori as simply not reasonable? Do any pairs match exactly and can they be a prior linked? That is, can you be sure of some decisions without having to put the record pairs through the algorithm? Doing so could save time without impacting accuracy to any great amount. Households with only one person present in both the decennial census and the post enumeration survey that agree or nearly agree on personal information might simply be declared to be matches. It also could have the advantage of helping algorithm performance. In mixture or latent class models, when you can use some labeled training data, it can effectively 'identify' the classes, thereby helping with interpretation and monitoring algorithm performance.
- How sensitive are results to choices (number of levels) for string comparators?
- The author consider two cases: mixture model parameters are the same in all blocks or they vary by block. Given that blocks are so small in this application, an alternative part way between these two extremes might be considered. Such as model could possibly be called a mixture model for sets of latent class parameters. Some neighborhoods might have high probabilities among matches of agreeing on matching fields. Other neighborhoods might tend to have lower probabilities. It might be reasonable to have one set of latent class model parameters in some neighborhoods, but other parameters in a different set of neighborhoods.

As mentioned previously, the authors' method is used to identify hard to link cells. Here are a couple of questions related to that goal.

- Only more data will help cases that are hard to link. Can a third data set (CARRA's population file) help in this regard? That is, can a third file with population information be linked to one input file or to both input files to add more information to the information available for judging match versus non match?
- What is important? Do you need to find actual links for counting the population, or do you need to be right 'on average'? In some cells with a lot of missing information, it is very hard to determine who is a match, but the proportion matches might be somewhat accurate.

5. Zach Kurtz, Smooth post-stratification in multiple capture-recapture

Zach Kurtz (2013) presented on smoothing post-stratification in multiple capture-recapture studies. In capture-recapture, one links individuals in 2 (or more) enumerations, determines "in" or "out" for each individual-enumeration, and employs a model to estimate the number in the "missed by all" cell. A third list can add information. See, for example, Zaslavsky and Wolfgang (1993). The innovation in this work is that probabilities of capture-recapture cells are made to depend on personal characteristics (sex, age in human context). Flexible modeling (kernel density estimates) is used in place of regular log linear models. The Bayesian Information Criterion (BIC) is used for model selection.

In record linkage context, all the variables used for matching/comparison might enter into models, so the log linear models might be quite involved. Perhaps there is a connection to calibration weighting in surveys: instead of simple raking survey weight adjustment or extensive post stratification, one can consider regression modeling more generally. See for example Gelman (2007).

The application considered in this paper is the American Bird Species Census for 2009-2011 which records presence/absence of species by year. Three questions come to mind.

- Can one make use of the number of sightings by species to get at probability of a sighting? A bird species with numerous sightings must have a higher probability of being present, whereas a bird with only a single sighting might have been missed had it not been seen that one time.
- Does the Census record sighting by region or state? If so, then stratifying by such a factor could improve models.
- The authors combine three Census years together in their capture-recapture estimation. How sensitive are results to using 2 or 4 years versus 3 years? There should be many ways to make comparisons. Do you get similar estimates using 2009 and 2011 as when you use all three years?

6. Summary

The authors of the papers in this session have produced innovative, interesting, and useful work. I look forward to seeing the published papers. Some relevant (generic) questions can be raised concerning these and similar efforts.

- What limitations of data will be the hardest to overcome? Can improvements in data quality be made and will they help?
- What additional diagnostics can be created and checked to increase the level of comfort with models and results? Are there additional relevant, realistic, and convincing simulations that can support the methods?
- What will convince policy makers that the methods are worthy of use in practice?

Here is an overall summary.

- This Invited session included high quality presentations on an important topics with diverse approaches and applications. The speakers and their colleagues deserve praise for their efforts.
- These approaches are relevant to ‘our times’: available files are large and contain a lot of information, but they are incomplete and (only) statistical models can extract more information.
- In addition to convincing ourselves that models are appropriate and helpful, we have to convince the end users and policy makers.

Acknowledgements

The author thanks the speakers for the talks and draft papers, the organizer Rebecca Steorts for putting the session together and the invitation to discuss the papers, the chair for running the session, and the audience for their attention and questions.

References

Gelman, A., (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2): 153-164.

Hall, R.C., Steorts, R.C., and Fienberg, S.E. (2013). Will the real Steve Fienberg please stand up? Getting to know a population from multiple incomplete files. Invited Paper Presentation, Would the Real Steve Fienberg Please Stand Up: Getting to Know a Population from Multiple Incomplete FiLes, Social Statistics Section, Session 572. 2013 Joint Statistical Meetings in Quebec, Montreal, Canada.

Kurtz, Z. (2013). Smooth post-stratification in multiple capture-recapture. Invited Paper Presentation, Would the Real Steve Fienberg Please Stand Up: Getting to Know a Population from Multiple Incomplete FiLes, Social Statistics Section, Session 572. 2013 Joint Statistical Meetings in Quebec, Montreal, Canada

Larsen, M.D. (2009). Record Linkage Modeling in Federal Statistical Databases. Federal Committee on Statistical Methodology 2009 Research Conference.

Mule, V.T., and Imel, L. (2013). Bayesian record linkage models for Census coverage measurement matching. Invited Paper Presentation, Would the Real Steve Fienberg Please Stand Up: Getting to Know a Population from Multiple Incomplete FiLes, Social Statistics Section, Session 572. 2013 Joint Statistical Meetings in Quebec, Montreal, Canada

Ventura, S., and Nugent, R. (2013). Deduplicating text records using clustering and aggregation of conditional classifies. Invited Paper Presentation, Would the Real Steve Fienberg Please Stand Up: Getting to Know a Population from Multiple Incomplete FiLes, Social Statistics Section, Session 572. 2013 Joint Statistical Meetings in Quebec, Montreal, Canada.

Zaslavsky, A.M., and Wolfgang, G.S. (1993). Triple system modeling of Census, Post-enumeration survey, and Administrative-list data. *Journal of Business and Economic Statistics*, 11(3): 279+