Bayesian Record Linkage Models for Census Coverage Measurement Matching

Vincent Thomas Mule, Jr. and Lynn Imel[1], U.S. Census Bureau

Abstract

This paper explores using record linkage Bayesian approaches for matching Census Coverage Measurement data to the Decennial Census. The methods proposed in Larsen (2009) will be examined and the methods will be extended to a) allow partial agreement of matching fields and b) account for comparisons of matching fields that are missing due to nonresponse. Additionally implementation of a one-to-one matching constraint will be discussed. The performance of the Bayesian approaches will be examined based on computer matching of the 2010 Census Coverage Measurement data.

Key Words: Record Linkage, Bayesian, Gibbs Sampling, Metropolis-Hastings

I. Introduction

The Census Bureau has a long history of researching and developing record linkage computer matching approaches. Jaro (1989) and Winkler and Thibaudau (1991) developed computer matching applications to support the 1990 Post-Enumeration Survey. Mule (2003) and Fay (2002, 2004) utilized computer matching approaches to produce the 2000 Accuracy and Coverage Evaluation Revision II estimates. Ikeda and Porter (2007, 2008) documented development of the computer matching approach for the 2010 Census Duplicate Person Identification and the Census Coverage Measurement (CCM) computer matching.

Larsen (2009) proposed Bayesian applications of record linkage along the lines of his earlier research based on the mixture model approach done by Fellegi and Sunter (1969). In this paper we apply some of the Larsen Bayesian applications to a subset of 2010 CCM computer matching data. Section II provides a brief description of the CCM data. Section III gives an overview of implementing Larsen's Bayesian approach that allows many-to-many matching and does not allow parameters to vary by block (blocks are a group of linked pairs that agree on at least one variable, such as agreement by phone number). Section IV shows how we modified Larsen's Bayesian approach to speed up processing time. Section V describes our implementation of Larsen's Bayesian approach that allows probabilities to vary by block. Section VI provides some results of applying a one-to-one restriction to the many-to-many links. Section VII provides future research ideas.

II. Census Coverage Measurement

The 2010 CCM program was the survey-based program to evaluate coverage of the 2010 Census to assist in planning improvements for future censuses, including 2020 and beyond. The program measured coverage of housing units and household population

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

(excluding group quarters and persons residing in group quarters) at the national level and various breakdowns such as demographic, geography and census operations. The coverage estimates included net coverage error.

The CCM sample size was approximately 170,000 housing units in the United States (excluding remote areas of Alaska). The CCM conducted interviews, an operation referred to as Person Interview (PI), at each housing unit in late summer of 2010. The interview collected demographic data and information to determine the person's residence on Census Day (April 1, 2010). The person data, PI data, were matched to the census enumerations. The entire CCM matching activity included computer matching and two stages of clerical review: one before and one after a data collection phase to help resolve a person's residence. The results of the CCM matching were the input to forming CCM coverage estimates. Our research dataset is from the CCM computer matching process.

The person computer matching process searched for matches between persons rostered at a CCM sample address and persons enumerated in the census. The process had two phases: the first consisted of two stages of record linkage, and the second was a modeling phase to decide which links were matches or possible matches. The first stage of matching consisted of running record linkage software, BigMatch, for various blocking and matching variables such as name, day of birth, age and sex. The BigMatch software produced match scores that were used later in processing to determine if linked pairs of records were matches or possible matches. For more on the BigMatch software, see Yancey (2007). This first stage of matching established a link between a CCM and census household. In the second stage, all person records in the linked households were rematched by the Center for Statistical Research & Methodology's One-to-One matching software. The software was run with the unique household identifier as the blocking variable. Processing with household as the blocking variable, ensured that for every household linked in the first stage, all of the CCM persons were matched against all of the census persons.

The modeling phase was implemented after the two stages of record linkage. The first step of this phase was calculating a link confidence metric for the results, CCM and census person links, of the second stage of record linkage. The metric was based on pre-defined rules that take into account numerous factors such as person characteristics. Headquarters staff reviewed the links after sorting by the confidence metric. The links were reviewed to pick cutoffs that determined potential matches. Next, links not identified as potential matches by the confidence metric and those that were only processed through the first stage of record linkage were evaluated by the match scores produced by BigMatch. Links with match scores above pre-determined cutoffs were identified as potential matches.

This paper focuses on applying Bayesian record linkage models to a subset of the input data for the second stage of record linkage. Our research data consist of the CCM and census households linked in the first stage of matching that reported the same phone number. In this subset, the CCM file has about 488,000 person records and the Census person file has around 492,000 person records. Using household as a blocking variable results in over 60,000 household blocking combinations and many-to-many matching results in over 760,000 links. Throughout the rest of this paper, we will examine how Bayesian alternatives perform as compared to computer matching.

III. Larsen Bayesian Record Linkage Models with Many-to-Many Matches and Parameters Not Varying by Block

Larsen (2009) describes Bayesian approaches to record linkage that link records from two files (A and B). The linked pair of records from files A and B are referred to as (a,b). Each pair of linked records has a comparison vector ($\gamma(\mathrm{a},\mathrm{b})$) that indicates an agreement pattern for K comparison fields. For a linked pair (a,b) the indicator of match status is defined as I(a,b) = 1 for match and I(a,b) = 0 for nonmatch.

Larsen approaches the problem based on two latent classes: matches and nonmatches. While the paper mentions the possible extension to three classes, we implemented the two latent classes. Since we were linking persons between housing units with the same phone number there was no benefit of a third latent class. Additionally, like other record linkage applications, Larsen makes the conditional independence assumption of the comparison variables. While interactions could be allowed, we make that assumption as well.

In this section, we will lay out the methodology described in section 3.1, *Bayesian Approach to Latent Class Record Linkage Models*, of his 2009 paper and our implementation of it on the subset of CCM data that we are using for this research.

As laid out in Larsen, the posterior distributions of the parameters were calculated using Gibbs sampling. The first step of the process was selecting initial values for the unknown parameters (initial parameters based on previous survey matching results). Next, as seen in Larsen, the following algorithm was repeated until convergence.

1. For each linked pair of records, draw values for the indicator of match status independently from a Bernoulli distribution. The probability of match is

$$\frac{p_m\, Pr(\gamma(a,b)|\, M)}{p_m\, Pr(\gamma(a,b)|\, M)+p_u\, Pr(\gamma(a,b)|\, U)} \qquad (1)$$

where, p_m = probability of match given match status indicator

p_u = probability of nonmatch given match status indicator

$Pr(\gamma(a,b)|\, M)$ = probability of observing pattern (comparison vector) among the matches

$Pr(\gamma(a,b)|\, U)$ = probability of observing pattern (comparison vector) among the nonmatches

2. Draw a probability of match given match status indicator from a Beta distribution. The probability of nonmatch is equal to 1 minus the probability of match.

$$p_m|I \sim Beta\left(\alpha_M + \textstyle\sum_{(a,b)} I(a,b), \beta_M + \sum_{(a,b)}\left(1 - I(a,b)\right)\right) \qquad (2)$$

3. For every k^{th} comparison field, draw* the probability of observing the comparison vector given the match probability and match status indicator.

$$Pr(\gamma_k(a,b) = 1|M, I) \sim$$

$$Beta\left(\alpha_{Mk} + \textstyle\sum_{(a,b)} I_{ab}\gamma_k(a,b),\ \beta_{Mk} + \textstyle\sum_{(a,b)} I_{ab}\left(1 - \gamma_k(a,b)\right)\right) \tag{3}$$

4. For every k^{th} comparison field, draw* the probability of observing the comparison vector given the nonmatch probability and the match status indicator.

$$Pr(\gamma_k(a,b) = 1|U, I) \sim$$

$$Beta\left(\alpha_{Uk} + \textstyle\sum_{(a,b)}(1 - I_{ab})\gamma_k(a,b),\ \beta_{Uk} + \textstyle\sum_{(a,b)}(1 - I_{ab})\left(1 - \gamma_k(a,b)\right)\right) \tag{4}$$

*As suggested in Larsen (2009), we specified prior distributions as two Dirichlet distributions.

We used a version of BigMatch that was developed for the 2000 Further Study of Person Duplication by the CSRM Record Linkage staff in our analysis to get the initial many-to-many linked pairs of records. This version of BigMatch provided the agreement or disagreement score for each individual comparison field (first name, last name, middle initial, month of birth, day of birth, age and sex). For first and last name, we wanted the comparison vector to account for more than just agreement or disagreement so the individual score allowed us to form five comparison levels for first and last name: exact agreement, strong partial agreement, weak partial agreement, disagree, and missing. For the remaining five comparison variables, we formed three levels: agree, disagree or missing. For any of the matching variables, missing was assigned if either or both of the records had a missing value.

We compared the results of our Bayesian implementation to the CCM computer matching results. Our implementation of the Bayesian approach yielded an estimate of 199,112 matches and the CCM computer matching identified 203,196 matches (matches refers to records identified as matches or possible matches, only a few of the records were possible matches) in these cases. The Bayesian estimate is based on the independent Bernoulli draw of matches taken for each iteration (after burnin) of the algorithm described above. (We ran the algorithm a total of 1,100 times with the first 100 iterations as burnin.)

To examine the differences, we identified the comparison vector combinations of the seven comparison fields that had differences of 100 or more matches as compared to the CCM. The Bayesian matches in the table are an estimate based on the number of links and the probability of being a match from the iterations (after burnin) of the algorithm. Table 1 shows these 17 comparison vector combinations.

Table 1: Comparison Vectors with 100+ Differences in Number of Matches

First Name	Last Name	Middle Initial	Month of Birth	Day of Birth	Age	Sex	# of Links	Mean Probability of Match	Matches		
									Bayesian Using Mean	CCM Computer Matching	Difference (CCM-Bayesian)
Exact	Exact	Missing	Disagree	Missing	Missing	Agree	1,167	0.230	268	1,080	812
Exact	Exact	Missing	Missing	Missing	Missing	Agree	1,488	0.728	1,084	1,428	344
Exact	Exact	Missing	Missing	Missing	Missing	Missing	395	0.410	162	373	211
Disagree	Exact	Missing	Missing	Missing	Missing	Agree	1,662	0.003	4	194	190
Exact	Exact	Missing	Disagree	Disagree	Disagree	Agree	1,739	0.001	2	184	182
Exact	Exact	Agree	Disagree	Disagree	Disagree	Agree	1,974	0.016	31	180	149
Exact	Disagree	Missing	Disagree	Missing	Missing	Agree	189	0.025	5	144	139
Disagree	Exact	Missing	Disagree	Missing	Missing	Agree	2,513	0.000	1	115	114
Exact	Exact	Missing	Disagree	Agree	Disagree	Agree	432	0.499	216	328	112
Exact	Exact	Missing	Disagree	Disagree	Agree	Agree	296	0.481	142	249	107
Exact	Disagree	Missing	Missing	Missing	Missing	Agree	164	0.185	30	136	106
Exact	Exact	Agree	Disagree	Missing	Missing	Agree	600	0.769	462	565	103
Disagree	Exact	Missing	Agree	Agree	Agree	Disagree	188	0.855	161	52	-109
Exact	Exact	Agree	Disagree	Agree	Disagree	Agree	570	0.918	523	387	-136
Disagree	Exact	Agree	Agree	Agree	Agree	Agree	1,725	1.000	1,725	1,577	-148
Disagree	Exact	Missing	Agree	Agree	Agree	Agree	2,483	0.998	2,478	2,197	-281
Disagree	Exact	Disagree	Agree	Agree	Agree	Agree	1,050	0.974	1,022	531	-491

Table 1 shows 12 of the 17 vectors are instances where CCM has more computer matches or possible matches than the Bayesian approach. When there are more CCM computer matches than estimated by our Bayesian approach, the mean probability of being a match falls into two groups. There are seven rows where the mean probability of match is between 0.185 and 0.769. The first row shows an instance where the links have exact agreement on first name, last name and sex but there is disagreement for month of birth and missing information for the remaining variables. The average match probability given this agreement pattern is 0.230. If you were to make an independent determination of these cases using this probability, the result would be about 20 percent of them being matches. Mule (2003) showed that month and day of birth are powerful discriminating variables for census record linkage applications when matching people within the same household. Since day of birth is missing and month of birth disagrees, it is not surprising that this comparison vector received a lower probability. In examining the final production matching results (after both stages of clerical review), 1,030 of the CCM records with this agreement pattern were confirmed matches, and we saw that these links were in housing units that contained links with high match probabilities. For the remaining 5 rows where there are more CCM computer matches than Bayesian links, the probability of being a match varies from 0.000 to 0.025. These are instances where either the first name disagrees or the month of birth, day of birth and age information either disagrees or is missing.

The bottom of Table 1 shows the five instances where the Bayesian approach estimates 100 or more matches than the CCM computer matching. The Bayesian mean probability of being a match for those instances ranges from 0.855 to 1.00. These vectors usually disagree on first name while agreeing on month of birth, day of birth, age and sex. We focus on the last row in the table that had the largest difference of 491 matches. In examining the final production matching results, we found that 968 of the CCM records are matches. The additional 437 matches were made through the clerical review.

IV. Binomial Application of Larsen Approach Not Varying By Block

Larsen (2009) lays out a Gibbs sampling approach that simulates the posterior distribution of the parameter by sampling from alternating conditional distributions. As shown in the previous section, the first part of the iteration was to sample whether a link was a match from a Bernoulli distribution with a probability determined by formula (1).

In our research scenario, each iteration required over 760,000 independent Bernoulli draws to determine match status. If we were to use this type of technique in future census applications the number of draws would be substantially larger. Using CCM as an example, we would be matching 400,000 sample records to 300 million census records. We want to look for ways to improve processing speed.

One way to speed up processing is to use a Binomial approach instead of a Bernoulli. In our matching, we have seven comparison variables. First name and last name comparisons have 5 levels of agreement. The remaining five comparison variables (middle initial, month of birth, day of birth, age and sex) have 3 levels of comparison that indicate agree, disagree or missing. Based on our matching comparison outcome space, there are 6,075 unique combinations (5×5×3×3×3×3×3). Instead of drawing individual Bernoulli outcomes, we used an algorithm based on the summarized dataset by counting the number of links for each of the 6,075 combinations. This allowed us to process 6,075 records instead of 760,000 plus records through each iteration. By using a Binomial approach, we drew the number of matches for each of the unique combinations based on the number of links and the probability of being a match for that iteration.

As expected, our research showed similar convergence and results for the Binomial and Bernoulli applications. In the future, we can use the Binomial application for summarized data when dealing with large sized files.

V. Larsen Hierarchical Bayesian Models Where Parameters Vary By Block

Another technique presented in Larsen (2009) is a hierarchical Bayesian model that allows probabilities to vary by block. Similar to the algorithm that does not allow parameters to vary by block it is an iterative process but the processing is very different because of the steps needed to incorporate specifying parameters by each s^{th} block. In this section, we will lay out the methodology presented by Larsen that allows probabilities to vary by block and our implementation of it on the subset of CCM data used in our research.

The algorithm uses Gibbs sampling to simulate the posterior distribution of the parameters and the Metropolis-Hastings algorithm to draw values of the hyperparameters. The first step of this process was choosing initial, unknown parameters and a match status indicator. Initial parameters were set lower than the specified values used in the Bayesian approach presented in Section III, and the assignment of match status indicator was based on iterations of that approach. After setting the initial values, we implemented the following steps, as seen in Larsen, by block *s* until convergence:

1. Draw a probability of match given match status indicator.

$$p_M | I, \alpha_M, \beta_M \sim Beta\left(\alpha_M + \sum_{(a,b)} I_{ab}\ ,\ \beta_M + n_a n_b - \sum_{(a,b)} I_{ab}\right) \qquad (5)$$

$$\textit{set constraint: } p_M \le \frac{min(n_a, n_b)}{(n_a n_b)}$$

where, n_a *= # records from file A in block, s*
n_b *= # records from file B in block, s*

2. For *k*th comparison field, draw a probability of match given match status indicator.

$$p_{Mk} | I, \gamma, \alpha_{Mk}, \beta_{Mk} \sim$$

$$Beta\left(\alpha_{Mk} + \sum_{(a,b)} I_{ab}\gamma_k(a,b),\ \beta_{Mk} + \sum_{(a,b)} I_{ab}\left(1 - \gamma_k(a,b)\right)\right) \qquad (6)$$

3. For *k*th comparison field, draw a probability of nonmatch given match status indicator.

$$p_{Uk} | I, \gamma, \alpha_{Uk}, \beta_{Uk} \sim$$

$$Beta\left(\alpha_{Uk} + \sum_{(a,b)}(1 - I_{ab})\gamma_k(a,b),\ \beta_{Uk} + \sum_{(a,b)}(1 - I_{ab})\left(1 - \gamma_k(a,b)\right)\right) \quad (7)$$

4. Draw values of hyperparameters from the Metropolis-Hastings algorithm described in Larsen (2009) Appendix A, *Metropolis-Hastings Sampling Steps for the Hierarchical Record Linkage Model*. Run the algorithm separately for the probability of match, and each comparison field matched and nonmatched.

5. For each linked pair, draw a match status indicator from a Bernoulli distribution with the match probability

$$\frac{p_m\, Pr(\gamma(a,b)|\, M)}{p_m\, Pr(\gamma(a,b)|\, M) + p_u\, Pr(\gamma(a,b)|\, U)} \qquad (8)$$

where, p_m = for block *s*, probability of match given match status indicator

p_u = for block *s*, probability of nonmatch given match status indicator

$Pr(\gamma(a,b)|\, M)$ = for block *s*, probability of observing pattern (comparison vector) among the matches

$Pr(\gamma(a,b)|\, U)$ = for block *s*, probability of observing pattern (comparison vector) among the nonmatches

For our research data, we implemented this algorithm using two levels of agreement: agree or disagree (includes missing). We removed links where there was only one-person link between the household from the CCM research data used in the previous sections.

The additional steps of processing each block and drawing values of the hyperparameters contributed to the processing time jumping from minutes to hours.

We compared the number of matches from the Bayesian application to the CCM computer matching results. Our implementation of the Bayesian approach yielded 185,536 matches compared to 186,040 matches or possible matches from CCM computer matching. The Bayesian estimate is based on the independent Bernoulli draw of matches taken for each iteration (after burnin) of the algorithm. Similar to the earlier work, we identified the comparison vectors with differences of 100 or more matches. Twenty-six vectors had differences over 100. CCM computer matches exceed the Bayesian estimate for 11 of the 26 vectors. For all of these vectors, month of birth and day of birth disagree. For the vectors where the Bayesian estimate of matches exceeds CCM, the majority of agreement patterns include first name disagreement and agreement for last name and sex.

This algorithm is in the initial stages of development. To use this algorithm we face challenges such as how to deal with a large number of blocks with only a few links and setting initial parameters. The results presented reflect how we dealt with the challenge of setting parameters. For initial parameter values, we set them to similar levels used in the Bayesian approach that does not allow parameters to vary by block; we lowered the initial values in subsequent runs to make them less influential. Since there are only a few links per block in this application, allowing the parameters to vary by block is more sensitive to the initial values selected. We are exploring ways to deal with the large number of blocks with only a few links, such as forming larger groups of the blocks.

VI. Implementing One-to-One Matching

Based on some of the results in the previous section, we proceeded to implement a one-to-one matching constraint onto the links. Larsen (2009) provided two possible ways of doing this. One, which was implemented in our analysis, was to feed the links into a linear sum assignment. The second was to modify likelihoods to account for this.

We used the SAS Proc Assign procedure to select the one-to-one links in this analysis. Proc Assign uses a variant of the out-of-kilter algorithm where the probability of being a match was used as the maximization variable. While this approach is not as optimal as the Hungarian algorithm or other linear sum assignment approaches, it was readily available and had a very high overlap with the same assignments made out of the CCM computer matching one-to-one assignment.

In this section, we will go over applying the one-to-one matching constraint to the Bayesian algorithm that does not allow the probabilities to vary by block shown in Section III. Applying the linear sum assignment by block to the research data in Section III resulted in just over 205,000 links. We removed links if the CCM record was in multiple blocks to assess the one-to-one matching assignment. The PI collected information for people at the housing unit at the time of the interview and for people living at the unit on Census Day. In some instances, the household roster differed for the two reference periods. For instance, a housing unit could have people who resided there at both reference periods and have people who lived there at only one of the reference days. An example is someone moving out over the summer. These situations required special treatment in the CCM matching. To make our analysis simpler, we excluded any of these types of situations.

After removing links with CCM records in multiple blocks, we had just under 198,000 links to review. Over 98% of the CCM records in the review links were in links identified as potential matches by computer matching. Fewer than 225 of these CCM records were matched to different census records. In addition, we followed the review links through clerical review. Just over 183,000 of the CCM records in the review links were matches in clerical review and over 98% of those records were matched to the same census record. In looking at the many-to-many results before and then the one-to-one results after, we see many times where links with probabilities less than 0.01 were not selected. While this is expected, it was another good point to see about the implementation of this algorithm on our data.

VII. Future Work

We have shown our initial research into applying a few of the Bayesian record linkage approaches described in Larsen (2009). The approaches yield estimates of matches that are similar to the match counts generated from CCM computer matching. When looking at specific patterns of the comparison fields, we see differences of over 100 matches.

Future work will focus on the Bayesian approach that allows probabilities to vary by block and one-to-one assignment algorithms. For the Bayesian approach that allows probabilities to vary by block, our focus is on expanding agreement levels, block definitions, and process improvements. We will focus on expanding the agreement levels because in general our survey data is at least 3 levels (agree/disagree/missing). Additionally, we will look into regrouping or redefining blocking passes because using phone number to define blocking passes yielded over 15,000 blocks with only one link. Our work will also focus on programming efficiency because when we allow probabilities to vary by block our processing time went from minutes to hours. We want to investigate using the Binomial approach summarized in Section IV to see if we can improve processing efficiency.

We want to investigate one-to-one assignment algorithms and how to classify the links selected from an algorithm as matches or nonmatches. The one-to-one research will continue because our initial results yielded a few many-to-many links (because of the way our data is structured). With our applications, usually we have to make a one-to-one determination of whether a CCM record matches to a census record. While there is duplication of the census that can result in the CCM having one-to-many matches to the census, it is usually just one-to-one.

Another future area of research is looking into using the Bayesian results to determine cutoffs that identify matches. Since we have 2000 and 2010 coverage measurement results that have computer and subsequent clerical assignments, we want to explore if there are ways to use Receiver Operator Curve (ROC) analysis to predict estimates of false matches and true matches not detected for different cutoffs based on sensitivity and specificity results observed in past CCM results.

VIII. References

Fay, R.E. (2002). “Probabilistic Models for Detecting Census Person Duplication,” *2002 Proceedings of the Joint Statistical Meetings on CD-ROM*, American Statistical Association, Alexandria, VA, 969-974.

Fay, R.E. (2004). “An Analysis of Person Duplication in Census 2000,” *2004 Proceedings of the Joint Statistical Meetings on CD-ROM*, American Statistical Association, Alexandria, VA, 3478-3485.

Fellegi, I.P. and Sunter, A.B. (1969). “A Theory for Record Linkage,” Journal of the American Statistical Association, 64, 1183-1210.

Ikeda, M. and Porter, E. (2007). “Initial Results from a Nationwide BigMatch Matching of 2000 Census Data,” SRD Statistics Research Report Series #2007-22, U.S. Census Bureau.

Ikeda, M. and Porter, E. (2008). “Additional Results from a Nationwide BigMatch Matching of 2000 Census Data,” SRD Statistics Research Report Series #2008-2, U.S. Census Bureau.

Jaro, M. (1989). “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida,” *Journal of the American Statistical Association*, 89, 414-420.

Larsen, M. (2009). “Record Linkage Modeling in Federal Statistical Database,” Federal Committee on Statistical Methodology 2009 Research Conference.

Mule, T. (2003). “Chapter 5 – Further Study of Person Duplication,” DSSD A.C.E. Revision II Memorandum Series PP-30r, U.S. Census Bureau.

Winkler, W. and Thibaudau, Y. (1991). “An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census,” SRD Technical Report Series #1991-9, U.S. Census Bureau.

Yancey, W. (2007). “BigMatch: A Program for Extracting Probable Matches from a Large File,” Research Report Series RRC2007/01, Statistical Research Division, U.S. Census Bureau.