CARRA Working Paper Series

Working Paper #2014-02

Estimating Record Linkage False Match Rate for the Person Identification Validation System

Mary Layne
U. S. Census Bureau

Deborah Wagner
U. S. Census Bureau

Cynthia Rothhaas
U. S. Census Bureau

Center for Administrative Records Research and Applications
U. S. Census Bureau
Washington, D. C. 20233

Paper Issued: July 1, 2014

Estimating Record Linkage False Match Rate for the Person Identification Validation System

July 1, 2014

Mary Layne
U. S. Census Bureau

Deborah Wagner
U. S. Census Bureau

Cynthia Rothhaas
U. S. Census Bureau

Abstract

The Census Bureau Person Identification Validation System (PVS) assigns unique person identifiers to federal, commercial, census, and survey data to facilitate linkages across files. PVS uses probabilistic matching to assign a unique Census Bureau identifier for each person. This paper presents a method to measure the false match rate in PVS following the approach of Belin and Rubin (1995).

The Belin and Rubin methodology requires truth data to estimate a mixture model. The parameters from the mixture model are used to obtain point estimates of the false match rate for each of the PVS search modules. The truth data requirement is satisfied by the unique access the Census Bureau has to high quality name, date of birth, address and Social Security (SSN) data. Truth data are quickly created for the Belin and Rubin model and do not involve a clerical review process. These truth data are used to create estimates for the Belin and Rubin parameters, making the approach more feasible. Both observed and modeled false match rates are computed for all search modules in federal administrative records data and commercial data.

TABLE OF CONTENTS

1. INTRODUCTION

The Census Bureau performs research using administrative records files to investigate methods to enhance Census Bureau statistical processes and products. Many projects at the Census Bureau involve matching persons across censuses, surveys, and federal data to enhance the understanding of participation in various federal programs. This work requires a method to ensure the same person is linked across multiple administrative files. By using the Person Identification Validation System (PVS) (Wagner and Layne, 2014) the Census Bureau establishes unique person and address identifiers.

The PVS uses probabilistic linking (Fellegi and Sunter, 1989) to match person data to reference files. Reference files are derived from the Social Security Administration (SSA) Numerical Identification file (Numident). The SSA Numident contains all transactions recorded against each Social Security Number (SSN) ever issued and is reformatted to create the Census Numident. The Census Numident reference file contains one record for each SSN, keeping all variants of date of birth (DOB) and name data in separate files. The Census Numident is enhanced with address information from administrative records to create another reference file, the GeoBase.

Through the PVS process, input person records that match the reference file are assigned unique person identifiers called protected identification keys (PIK). A PIK is an anonymous identifier as unique as a SSN. Once assigned, the PIK serves as a person linkage key across all files that have been processed using PVS. The PIK also serves as a person unduplication key within files.

When linked files are used for analysis, error in the linkage process can cause problems when estimating the relationship between variables on different files (Lahiri, Larson, 2005). Analysts need to be aware of both the error introduced by false matches and false non-matches to make adjustments in statistical methods.

This paper discusses the Center for Administrative Records Research and Applications' (CARRA) implementation of Belin and Rubin's (1995) method for estimating false match rates. Belin and Rubins's approach requires truth data[1]. We created truth data based on CARRA's unique access to high quality federal and SSN data by extracting the verified records from the PVS Verification[2] module. In other studies, truth data are produced via labor and time intensive clerical review, but here we used the Census Numident to automatically determine the "truth" of a match. This approach made the determination extremely fast and more feasible than clerical review. Using these truth data we were able to employ Belin and Rubin's methodology.

[1] Truth data are matched records that are labeled as either a true match or false match.

[2] The PVS Verification module performs a direct match to the Census Numident based on SSN, and confirms agreement of name and date of birth data.

2. BACKGROUND

The primary purpose of the PVS is to assign a unique PIK to records on a file to facilitate matching and unduplication. A file going through the PVS process begins with an initial edit to clean and standardize the name, date of birth, sex, and address of each record. Records then cascade through six search modules with different blocking and matching variables (Wagner and Layne, 2014).

Verification Module
If the input file has SSN data, it first goes through a verification process. The verification module matches the reported SSN from the incoming file to the Census Numident file, along with the Alternate Name and Alternate Date of Birth Numident files. If the SSN is located in the Census Numident, and the name and date of birth agree sufficiently, the SSN is considered verified and the PIK for that SSN is assigned. The SSN verification module is an exact match to SSN, so no match parameter file[3] is required for this step.

GeoSearch Module
The GeoSearch module searches for PIKs for incoming records that failed the verification module or are without reported SSNs or Individual Taxpayer Identification Numbers (ITINs) (assigned by the Internal Revenue Service). This module links records from the incoming file to the GeoBase through blocking passes[4] defined in the parameter file. The GeoBase Reference file contains addresses found in multiple years' federal files for SSNs in the Census Numident file, and includes data for ITINs.

The typical GeoSearch strategy starts by blocking records at the household level, the lowest level of geography, then broadens the geography for each successive pass and ends by blocking by the first three digits of the ZIP Code. The typical matching variables are first, middle, and last names; generational suffix; date of birth; gender and various address fields. For example, consider the input address 123 Main Street, Apartment A, Washington DC, 20001: the most stringent blocking would seek a match within the housing unit at Apt. A, and the most lenient blocking approach would seek a match within all addressees in ZIP Codes starting with 200. Matches are made based on sufficient agreement of the name and date of birth fields of the input file.

After the initial set of links is created in GeoSearch, a post-search program is run to determine which of the links are retained. A series of checks are performed: First the date of death information from the Numident is checked and links to deceased persons are dropped. Next a check is made for more than one SSN assigned to a source record. If more than one SSN is

---

[3] Each search module has its own parameter file. It includes the number of passes, and for each pass, the file details blocking keys, the matching variables, comparison type for the match, matching weights, and the cut-off value for determining whether a match has been made. The pass number determines how many times a match will be attempted, with various configurations and rules for the matching variables.

[4] Blocking divides the comparison space into manageable pieces. For example, suppose file A contains 300 million records and File B (the reference file) contains 1.2 billion records. The comparison space is 300,000,000*1,200,000,000. This is an unmanageable number because the comparison would take too long and probably wouldn't even be possible with the computer resources available to us. When using a5-digit ZIP blocking variable the data are cut into 5-digit ZIP Code chunks, and the search comparisons are done within these cuts.

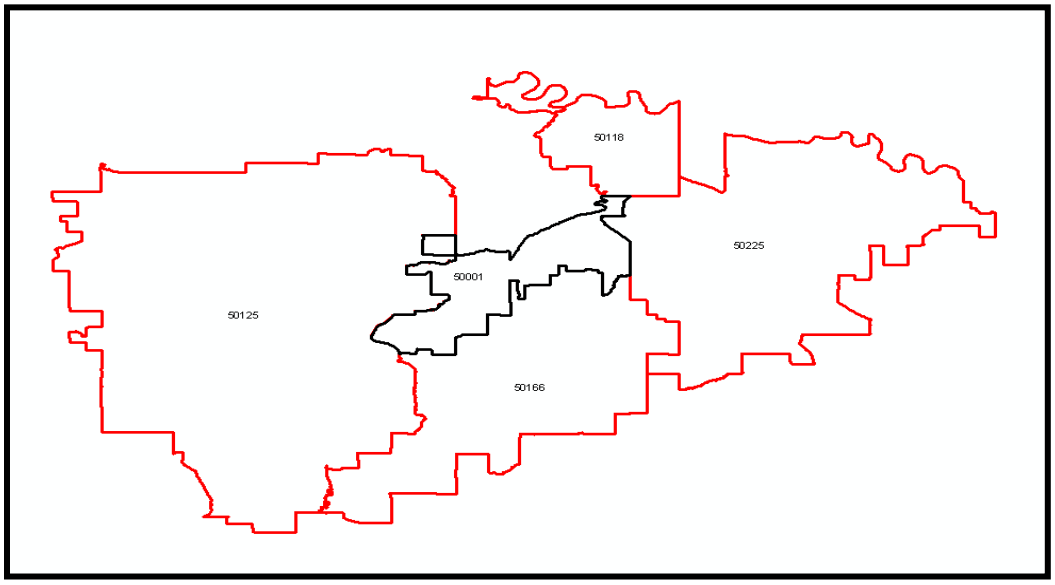
assigned, the best link is selected based on match weights. If no best SSN is determined, all SSNs assigned in the GeoSearch module are dropped and the input record cascades to the next module. A similar post-search program is run at the end of all search modules.

ZIP3 Adjacency Module

The ZIP3 Adjacency module searches reference files for records failing the Geosearch and incorporates the adjacency of neighboring areas with different ZIP3 values (Miller, Layne, Bouch, Smith, 2013). This can eliminate the bias of limiting the GeoSearch blocking strategy to exact match on ZIP3 and can find links when there are miscodes in the ZIP3 field.

Spatial adjacency is a geospatial term and means two polygons (in our case, ZIP3 codes) share a boundary (ESRI, 2012). Figure 1 provides an example of the notion of spatial adjacency. As shown in Figure 1, ZIP3s 501 and 502 are contiguous ZIP Codes to ZIP3 500. All persons with addresses in ZIP3 500 that failed the previous GeoSearch have the opportunity to be searched within the ZIP3 values of 501 and 502. If there were a typographical error in the first three digits of a ZIP Code, or the person had moved to a close-by ZIP3 and they were represented in the GeoBase at the new ZIP3, the ZIP3 module would find the link (assuming all other matching parameters are the same).

Figure 1. Spatial Adjacency



Movers Module (Prototyped only – not implemented in PVS)

The Movers module is appropriate for input files that combine individuals together into households. The module seeks multiple members of an input household in one address that may have moved together to another address. To be eligible for this module the household size must be greater than one. This module is being tested and was not used in the analysis for this paper.

Name Search Module
The Name Search module searches the reference files for records failing the previous modules. Only name, gender, and date of birth data are used in this search process. Name Search consists of multiple passes against the Numident Name Reference file, which contains all possible combinations of alternate names and alternate dates of birth for each SSN in the Census Numident file, and includes data for ITINs.

The typical Name Search blocking strategy starts with a strict first pass, blocking records by exact date of birth and parts of first and last names. Successive passes block on parts of the name and date of birth fields to allow for some name and date of birth variation. Matches are made when there is enough agreement on the matching variables.

As an example, consider the input record with first and last name "Seymour Henderson II" born on March 9, 1956: the strictest blocking pass searches for a match by using the NYSIIS code[5] for first (SAYNAR) and last (HANDARSAN) name and an exact match on the date of birth. The last, and least strict pass, blocks on the first two characters of the first and last name ("Se" and "He") and date of birth month (03) and year (1956).

Date of Birth (DOB) Search Module
The DOB Search module looks through the reference files for the records that fail the Name Search, using name, gender, and date of birth data. The module matches against a re-split version of the Numident Name Reference file, splitting the data based on month and day of birth.

There are typically four blocking passes in the DOB Search module. The first pass blocks records by first name in the incoming file to last name in the DOB Reference file and last name in incoming file to first name in the DOB Reference file. This strategy accounts for switching of first and last name in the incoming file. Subsequent passes block on various parts of first and last name and allow for some variation in date of birth. Matches are made when there is enough agreement on the matching variables.

Consider again "Seymour Henderson II" born on March 9, 1956. The first blocking pass searches for a match where first and last name have been switched in the input data. This pass looks for matches where the first name is "Henderson" and the last name is "Seymour." The last pass blocks on the first two characters of the first name ("Se") and the year of birth (1956).

[5] The New York State Identification and Intelligence System Phonetic Code, commonly known as NYSIIS, is a phonetic algorithm devised in 1970 as part of the New York State Identification and Intelligence System. This algorithm is designed to work specifically with American names. It creates indices for words based on their pronunciation. The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling.

Household Composition Search Module
The Household Composition Search module searches the reference files for records that have failed the DOB Search and preceding modules. To be eligible for this module, at least one person in the household of an unmatched person must have received a PIK. This module creates an eligible universe by selecting all not-found persons from the input data where at least one person in their household received a PIK. A reference file is created at run-time.

For example, consider an eligible three person household where two are given a PIK (Howard and Beth) and one is not (Kenny). We extract all of the records associated with addresses for Howard and Beth from the Geobase Reference file. This provides a list of all addresses associated with Howard and Beth and, potentially, Kenny. The resulting list is searched for Kenny's name and date of birth information from the input data and if found, Kenny then receives the PIK associated with the record. (Recall the Geobase Reference file is a file with multiple years of addresses found in federal data with PIKs.)

For this research, we modeled false match rate for the GeoSearch, ZIP3 Adjacency, Name Search and Date of Birth modules. Two modules were not included in the analysis: The Mover's module is currently not in production and the Household Composition module was recently implemented and is in the process of being vetted.

3. DATA DESCRIPTION

We required federal and commercial data to contain SSN to create truth data without employing clerical review. As described above, the input record SSNs were directly matched with the Census Numident.

3.1 Federal Data

We chose the Medicare Enrollment Database (MEDB) from the Centers for Medicare and Medicaid Services (CMS) for 2011 and 2012. These data contain Medicare enrollee data.

In addition, we used the Indian Health Service (IHS) Patient Registration File. The IHS Patient Registration File contains information on American Indians or Alaska Natives (AIAN) who participate in the IHS System. Spouses and children of AIANs not in this race group are eligible to receive these services as well. Each visit an individual makes to an IHS facility their insurance, demographic, and employment information is verified and updated, if needed. Therefore multiple records for an individual in the file may exist. Associated with each of these records is the transaction date of the record entry. We selected the most recent transaction date to avoid multiple records for a person in the verified data.

We chose the MEDB and IHS files because they both have high quality personal identifiers and represent the files with the highest verification rate (MEDB at 99.9%) and lowest verification rate (IHS at 96.7%) among federal administrative record sources. For both the MEDB and the IHS we performed analysis on two years of data (2011 and 2012) to monitor the consistency of observed and modeled estimates.

3.2 Commercial Data

The U.S. Census Bureau acquired data files containing identifying information and demographic characteristics from commercial data vendors. Most of the files contained current information on address, name, race, Hispanic origin, age, and sex data. Few files contained historical data on the same variables.

We used the 2010 commercial data from two separate vendors, Vendor 1 and Vendor 2[6], who sold SSN data. Vendor 1 data were only processed through the GeoSearch, Name Search, and ZIP3 Adjacency Modules. The DOB Search module could not be attempted because Vendor 1 did not provide day of birth data and the DOB Search module requires full date of birth. For the DOB false match rate estimates we used verified Vendor 2 data, which had the full set of date of birth variables, to estimate false match in the DOB module. Vendor 2 data were not processed through any other search modules.

3.3 Analysis Data

False match rates were estimated for the following ***groupings*** of data:

- ***Class of data***
  - Two federal data sets for both 2011 and 2012 (four data sets total)
  - One 2010 commercial data set
    - ***Search module:*** GeoSearch, ZIP3 Adjacency, Name, and Date of Birth
      - ***Passes within modules in which the cut-off value was different***
        - GeoSearch passes 1-5 (same cut-off value for each pass), 6, 7, 8, and 9
        - ZIP3 Adjacency Search passes 1 and 2
        - Name Search passes 1-4 (same cut-off value for each pass)
        - Date of Birth Search passes 1-4 (same cut-off value for each pass), using Vendor 2 data.

We will use the word ***grouping*** for the remainder of the paper to refer to the data sets we analyzed as classified above. An example of a particular grouping is:

2011 Federal Data Set (Medicare Enrollment Database), GeoSearch module, passes 1-5

We stripped SSN from the PVS verified data and processed the records through each probabilistic search module in the PVS. For example, a Medicare record contained an individual's SSN, name, and date of birth. The Verification module of PVS searched the Medicare SSN in the Census Numident and confirmed values in the other identifiers, assigning a verified PIK. The SSN was then stripped, and the Medicare name, date of birth, and address data were processed through each of the search modules. When the name, date of birth, and address data matched a record in the reference file, a PIK was assigned to the record.

[6] As part of the requirements of receiving commercial data, we are unable to reveal the sources of individual datasets.

We compared the probabilistically assigned PIK to the PIK in the verified data. Data were labeled as false matches when the PIK assigned in a probabilistic module did not match the PIK assigned in PVS Verification module. Appendix A presents the blocking and matching variables we used in each search module.

We were thus able to obtain observed estimates of the false-match rate and create truth data for modeling. Our model was based on the work of Belin and Rubin (1995), which used labeled data and estimated a two-class mixture model and determined the false match component.

4. STATISTICAL METHODLOGY

Two steps were used to develop the PVS false match rate. The first step was an analysis of observed false matches in truth data. The second step was to use the Belin and Rubin (1995) methodology to model false match rates using truth data.

4.1 Observed False Matches

To obtain observed false matches in the truth data, we first extracted the verified records from production runs of the Verification module. We then processed these verified records through each search module, ignoring the input SSN, and obtained the probabilistic PIK assignment. We compared the PIK found in the verification process with the PIK assigned in each search module. When the probabilistic PIK did not match the verified PIK, the record was labeled as a false match. This yielded observed false matches and labeled truth data.

We used these labeled truth data for our implementation of the Belin and Rubin model. The observed false matches were used to compare with the results obtained from the mixture model.

4.2 Belin and Rubin Methodology

We developed a version of Belin and Rubin's false match rate model. The model makes use of the weights from each search module, in combination with truth data, to determine the false match rate for a given cut-off value for that module. The key to the approach is viewing the distribution of observed match weights as a mixture of the distribution of weights for true matches and the distribution of weights for false matches.

Thomas Belin used Fortran to create a program for his model, using data from the 1990 Census Post Enumeration Survey (PES) as the truth deck for the 1990 Census responses. We used the SAS® IML language to follow Belin's Fortran logic closely. We used SAS® IML because of its ease of use with matrices and statistical functions and ability to directly read PVS data, which are SAS datasets. We made extensive use of the SAS macro language and shell scripts to generalize the programs for a production environment.

The modeling involved three steps. First, we estimated parameters needed to transform the match weights into a normal distribution. A separate set of transformation parameters were estimated for each grouping of data.

Applying the transformation parameters to the weights, we next fit a two group (true and false matches) mixture model to the transformed-normal distribution of weights. We used maximum likelihood estimation (MLE) to obtain estimates of the mixing proportions. The third step followed Belin and Rubin's method for obtaining a point estimate of the false match rate at the cut-off values used in the search module.

4.2.1 Estimation of Box-Cox Parameters to Transform Match Weights into Normal Distribution

The Belin and Rubin model requires the true and false distributions of the match weights to be normal. Our match weights are not normally distributed, so we used a Box-Cox power transformation of the weights, as suggested in the Belin and Rubin paper. This transformation scales the data by a function of observed data, which creates a rank-preserving transformation.

Using the labeled data for each grouping, we estimated the following equation separately for data labeled as true and data labeled as false.

$$\psi(w_i;\ \gamma, \omega) = \frac{w_i^{\gamma} - 1}{\gamma^{\omega^{\gamma-1}}} \qquad if\ \gamma \neq 0$$
$$= \omega \log(w_i) \qquad if\ \gamma = 0$$

Where:

w_i - Is the composite weight, which is the sum of match weights for each comparison variable in a pass. The composite weight is used to determine if the record is a match or not, based on the cut-off value[7] for the module and pass.

γ – Is the power parameter, which scales larger values more than smaller ones, resulting in a distribution which is more symmetric and closer to normal.

ω – Is the geomean of the composite weights (w_i), which is used to scale transformations such that likelihoods for different values of γ are directly comparable. Also, the size and range of transformed values depend highly on γ if not scaled by the geometric mean.

We estimated values of γ^F , γ^T , ω^F , ω^T . *T* and *F* superscripts indicate the calculations are made for the true and false distributions of weights separately.

We created a variance ratio (*V*), defined as $\frac{\sigma_F^2}{\sigma_T^2}$ for each grouping of data. Belin and Rubin used these as fixed estimates in the mixture model. The variance ratio (*V)* was found by taking the ratio of variances of transformed weights ψ_{Fi} and $\psi_{Ti}, i = 1, 2, \ldots . n.$

[7] If the sum of all match comparison weights is above this value, the two records are classified as a match.

4.2.2 Mixture Model

Next, we used an Expectation-Maximization (EM) algorithm to obtain Maximum Likelihood Estimates (MLE) (McLachlan and Peel, 2000) of the parameter estimates for the mixture of transformed normal distributions, ψ_{Fi} and $\psi_{Ti}, i = 1, 2, \ldots . n$.

The mixture model is defined as:

$$W_i \mid \theta, \{Z_i = 1\} \sim \text{Transformed } N(\mu_F, \sigma_F^2, \gamma_{F,}, \omega_F)$$
$$W_i \mid \theta, \{Z_i = 0\} \sim \text{Transformed } N(\mu_T, \sigma_T^2, \gamma_{T,}, \omega_T)$$
$$Z_i|\theta \sim \text{Bernoulli } (\lambda)$$

We used the EM algorithm to estimate θ $\{\mu_F, \mu_T, \sigma_F^2, \lambda\}$. Using Bayes Theorem, the *E* step at the (t+1)st iteration involved calculating:

$$Z_i^{(t+1)} = E(Z_i|W_i, \ldots, W_{n;}\ \phi^t, \phi\lambda^t)$$

where ***t*** is the iteration number.

We solved for:

$$Z_i^{(t+1)} = \frac{\lambda^{(t)} f_F\left(W_i \middle| \mu_F^{(t)}, \sigma_F^{2(t)}\right)}{\lambda^{(t)} f_F\left(W_i \middle| \mu_F^{(t)}, \sigma_F^{2(t)}\right) + (1-\lambda^{(t)}) f_T\left(W_i \middle| \mu_T^{(t)}, \sigma_T^{2(t)}\right)}$$

where,

$$f_F\left(W_i \middle| \mu_F^{(t)}, \sigma_F^{2(t)}\right) = \frac{1}{\sigma_F\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left\{\frac{\psi_i - \mu_F}{\sigma_F}\right\}^2\right]\left(\frac{(|W_i|+1)^{2\delta_0(W_i)-1}}{\omega_F}\right)^{\gamma_F - 1}$$

The $\psi_i's$ are transformed weights (W_i) corresponding to false match Box-Cox parameters.

We also calculated $f_T\left(W_i \middle| \mu_T^{(t)}, \sigma_T^{2(t)}\right)$ by replacing the subscript *F* with *T* to obtain the normal probability density for the distribution of weights from matches confirmed as true using truth data.

We calculated the estimated parameters as follows.

Mixing parameter, λ:

$$\lambda^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} Z_i^{(t+1)}$$

Variance components:

$$\sigma_F^{2(t+1)} = \frac{1}{n}\left[\sum_{i=1}^{n} Z_i^{(t+1)}\left\{\psi_i(W_i;\gamma_F,\omega_F) - \mu_F^{(t+1)}\right\}^2\right.$$

$$\left. + V\left(1 - Z_i^{(t+1)}\right)\left\{\psi_i(W_i;\gamma_T,\omega_T) - \mu_T^{(t+1)}\right\}^2\right]$$

$$\sigma_T^{2(t+1)} = \frac{\sigma_F^{2(t+1)}}{V}$$

Means:

$$\mu_F^{(t+1)} = \frac{\sum_{i=1}^{n} Z_i^{(t+1)}\, W_i}{\sum_{i=1}^{n} Z_i^{(t+1)}}$$

$$\mu_T^{(t+1)} = \frac{\sum_{i=1}^{n}\left(1 - \sum_{i=1}^{n} Z_i^{(t+1)}\right) W_i}{\sum_{i=1}^{n}\left(1 - \sum_{i=1}^{n} Z_i^{(t+1)}\right)}$$

The above equations were computed until convergence was achieved. In our modeling all of the search modules converged, except ZIP3, Pass 2.

4.2.3 False Match Rate Estimation

Once λ is estimated, we calculated a point estimate of the false match rate, given a particular cut-off value as:

$$\varepsilon(C|\theta) = \frac{\lambda\left[1 - \Phi\left(\frac{\psi_F(C;\,\lambda_F,\omega_F) - \mu_F}{\sigma_F}\right)\right]}{\lambda\left[1 - \Phi\left(\frac{\psi_F(C;\,\lambda_F,\omega_F) - \mu_F}{\sigma_F}\right)\right] + (1 - \lambda)\left[1 - \Phi\left(\frac{\psi_T(C;\,\lambda_T,\omega_T) - \mu_T}{\sigma_T}\right)\right]}$$

where Φ is the cumulative normal distribution.

5. RESULTS

In this section, we discuss the observed false matches and false match rates from our application of the Belin and Rubin model. Our expectation was the federal data files would have fewer false matches than the commercial files for both observed and modeled approaches. We anticipated federal file information to be more accurate because in many instances benefits are dependent on the veracity of SSN, name, date of birth, and address. We also performed analysis on a second year of data, using the 2012 MEDB and IHS data, to verify consistency in our estimates. The 2012 estimates were almost identical to the 2011 results; therefore only tables for 2011 are included in this paper.

5.1 Observed False Matches

We obtained observed false matches by processing records from PVS Verified data, stripped of SSN, through each PVS search module. We compared the verified PIKs with the PIKs assigned in the search modules and computed the number of non matches. We computed an observed match error as a proportion of false matches to total records assigned a PIK and reported them as percentages.

In PVS production settings, records *cascade* through the PVS – and only records failing verification or a particular matching module proceed on to the next module. If a record has a SSN and is verified by comparison to the Numident, the record doesn't go to the search modules. The verified records we ran through each of the modules would not have normally gone through to search because they would have been verified by SSN. Additionally, SSN is a matching variable in each of the search modules, but this research did not include SSN as a matching variable. If a record is assigned a PIK from the reference file in any of the modules, no further searches are conducted. The Name and DOB modules are towards the bottom of the module order and were intended to catch records that failed other modules.

Tables 1 -3 display the results[8] of observed false match percentages for 2011 federal and 2010 commercial data. The top row in each table, *Total Verified*, is the number of records for each data set (MEDB, IHS, Vendor 1, and Vendor 2[9]) that were verified via the PVS Verification process. These data were run through each of the search modules independently.

In each of the tables, the row of totals under each search module represents the number of Verification observations assigned a PIK in the search module. The **Total Verified** and module **Totals** differ because not all observations in the Verification data were assigned a PIK in the search modules. The lower total number verified for the ZIP3 searches are a function of the intent of the module. Recall the ZIP3 module searches ZIP3s adjacent to the ZIP3 on input file. It therefore ignores the ZIP3 the input record is in, where most matches are made.

[8] We excluded the seventh and eighth passes of GeoSearch in the tables because of small numbers in these cells.

[9] For 2010 commercial data the Vendor 2 commercial data were used only for the DOB module. The Vendor 1 data were used for all other modules.

Table 1 presents the results for 2011 MEDB data. The observed error for GeoSearch passes 1-4 is effectively zero (.004%). GeoSearch passes 5-6 have a high observed error percentage (10.828%), but the number of observations in these passes is very low. The total error for the entire GeoSearch module is a mere .005%. ZIP3 Pass 2 has the next highest match rate at (1.177%), followed by Name Search (.262%) and DOB Search (.177%).

Table 1. Observed Error – 2011 MEDB

	2011 MEDB			
	Number of Observations	**Search PIK Matches Verified PIK**	**Search PIK Doesn't Match Verified PIK**	**% Observed False Matches**
Total Verified	**53,058,202**			
GeoSearch				
Passes 1-4	52,186,950	52,184,681	2,269	0.004%
Passes 5-6	157	140	17	10.828%
Pass 9	219,874	219,575	299	0.136%
TOTAL	**52,406,981**		**2,585**	**0.005%**
Zip3 Spatial Adjacency				
Pass 1	11,737	11,735	2	0.017%
Pass 2	5,159,187	5,098,480	60,707	1.177%
TOTAL	**5,170,924**		**60,709**	**1.174%**
NameSearch				
Passes 1-4	49,374,794	49,245,314	129,480	0.262%
DOBSearch				
Passes 1-4	50,327,034	50,237,940	89,094	0.177%

Source: 2011 MEDB.

Table 2 details observed error for the 2011 IHS data. The table shows that the lowest observed error is in GeoSearch passes 1-4 (.046%) and the highest is found in GeoSearch pass 9 (.892%)[10].

Table 2. Observed Error – 2011 IHS

	2011 IHS			
	Records	**Search PIK Matches Verified PIK**	**Search PIK Doesn't Match Verified PIK**	**% Observed False Matches**
Total Verified	**2,782,181**			
GeoSearch				
Passes 1-4	2,469,032	2,467,900	1,132	0.046%
Passes 5-6	204	197	7	3.431%
Pass 9	11,209	11,109	100	0.892%
TOTAL	**2,480,445**		**1,239**	**0.050%**
Zip3 Spatial Adjacency				
Pass 1	796	795	1	0.126%
Pass 2	557,909	555,438	2,471	0.443%
TOTAL	**558,705**		**2,472**	**0.442%**
NameSearch				
Passes 1-4	2,458,244	2,440,698	17,546	0.714%
DOBSearch				
Passes 1-4	2,543,906	2,533,936	9,970	0.392%

Source: 2011 IHS.

[10] We excluded GeoSearch pass 5-6 because of the small number of observations.

Table 3 shows observed error in the 2010 commercial data, where the highest errors are found in GeoSearch pass 9 (13.113%), Name Search (4.177%) and DOB Search (6.604%). The GeoSearch total module observed error is quite low (.185%). Vendor 2 data were processed for the Date of Birth module and no others.

Table 3. Observed Error – 2010 Commercial

	2010 Commercial			
	Records	**Search PIK Matches Verified PIK**	**Search PIK Doesn't Match Verified PIK**	**% Observed False Matches**
Total Verified	**Vendor 1: Total Verified = 210,587,934**			
GeoSearch				
Passes 1-4	161,984,045	161,747,967	236,078	0.146%
Passes 5-6	12,351,322	12,276,474	74,848	0.606%
Pass 9	87,033	75,620	11,413	13.113%
TOTAL	**174,422,400**		**322,339**	**0.185%**
Zip3 Spatial Adjacency				
Pass 1	7,412	7,402	10	0.135%
Pass 2	212,834	209,756	3,078	1.446%
TOTAL	**220,246**		**3,088**	**1.402%**
NameSearch				
Passes 1-4	98,862,854	94,733,395	4,129,459	4.177%
	Vendor 2: Total Verified = 179,860,081			
DOBSearch				
Passes 1-4	65,313,236	60,999,837	4,313,399	6.604%

Source: 2010 Vendor 1 and Vendor 2.

In Table 4, we compare the observed error rates for all three data sets. In general, observed error is lowest for the 2011 MEDB and highest for the 2010 commercial data. This is as expected and consistent with the high quality of MEDB data versus IHS (which is still high quality) and commercial data. Indeed, observed error increases steadily for Name Search (.262% in the MEDB, .714% IHS, and 4.177% in commercial data) and DOB Search (.177% in the MEDB, .392% IHS, and 6.604% in commercial data). This trend is also seen in the total GeoSearch error. The ZIP3 total observed error do not follow this pattern.

The higher observed error percentages in the 2010 commercial data for Name Search and DOB Search are deceptive. When the PVS is run normally, very few records cascade down to these searches, having been found in either Verification or previous modules. The false matches in these research data may have been more precisely matched in the GeoSearch or other modules.

Vendor 1 data were missing day of birth (but contained year and month), so these data were matched on a less restrictive pass of the Name module, where day of birth was allowed to be missing. Vendor 2 day of birth values were clumped at the first day of the month (13%), so this may have resulted in the relatively large error in the DOB Module. In addition, Vendor 2 had missing values for 50% of the DOB fields. These results suggest we may consider making the cut-off for Name and DOB searches higher.

Table 4. Comparison of Observed Error

	2011 MEDB	2011 IHS	2010 Commercial
	% Observed False Matches	% Observed False Matches	% Observed False Matches
GeoSearch			
Passes 1-4	0.004%	0.046%	0.146%
Passes 5-6	10.828%	3.431%	0.606%
Pass 9	0.136%	0.892%	13.113%
TOTAL	**0.005%**	**0.050%**	**0.185%**
Zip3 Spatial Adjacency			
Pass 1	0.017%	0.126%	0.135%
Pass 2	1.177%	0.443%	1.446%
TOTAL	**1.174%**	**0.442%**	**1.402%**
NameSearch			
Passes 1-4	0.262%	0.714%	4.177%
DOBSearch			
Passes 1-4	0.177%	0.392%	6.604%

Source: 2011 MEDB, 2011 IHS, and 2010 commercial Vendor 1, Vendor 2.

5.2 Modeled False Match Rates

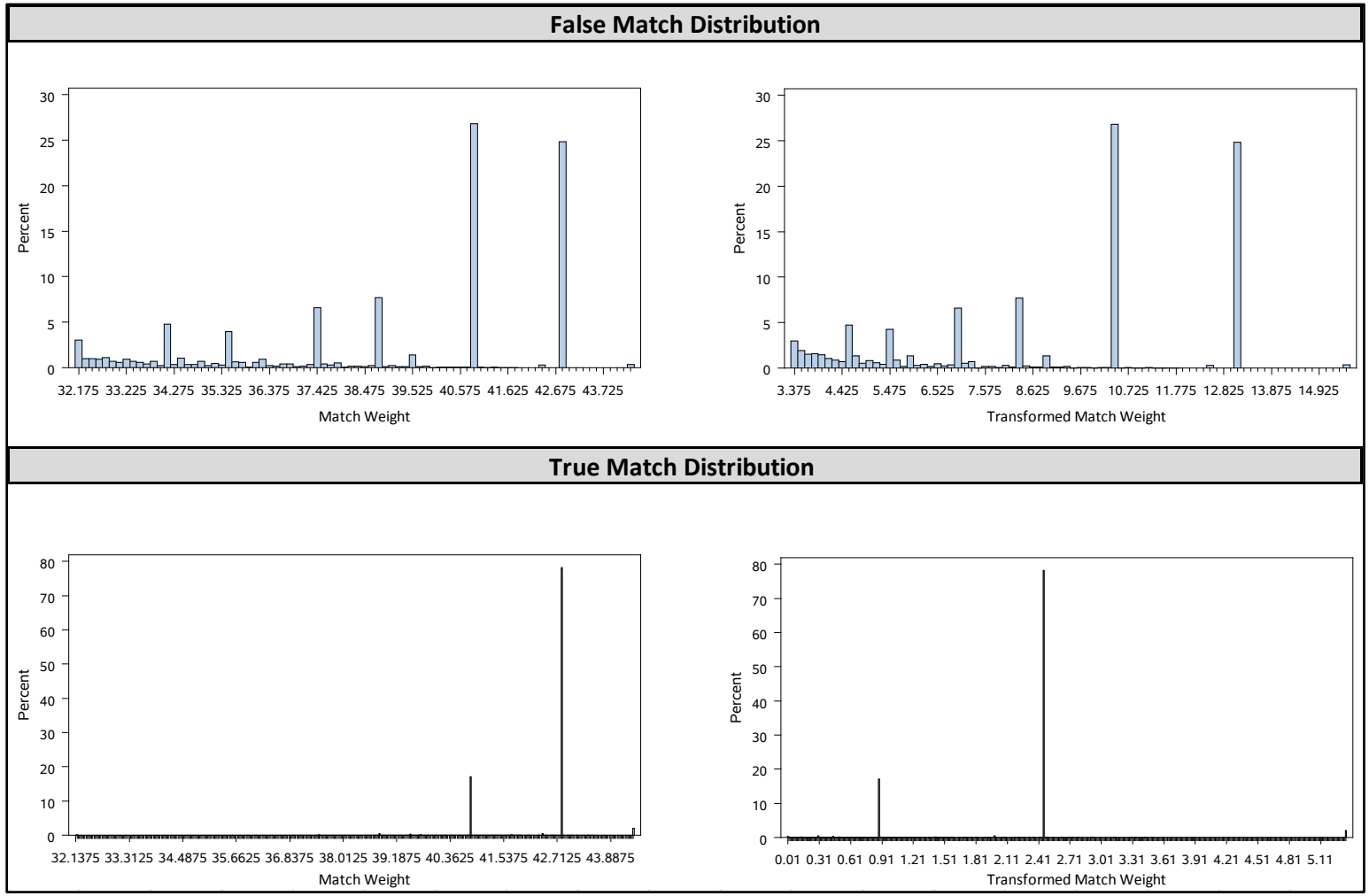
The Belin and Rubin false match rate model requires the true and false distribution of match weights to be normally distributed. Figures 2 through 4 display the false match distributions at the top and the true match distributions at the bottom. Figure 2 shows the match weight distributions for 2011 MEDB truth data for Name Search (pass 1-4). Figure 3 shows the match weight distributions for 2011 IHS truth data for GeoSearch (passes 1-4), and Figure 4 depicts the match weight distributions in 2010 commercial truth data for the DOB Search module. These particular graphs are shown to illustrate the different data types across the three PVS modules studied in this paper. The figures also present the weight distributions after a Box-Cox transformation.

The figures are very similar across all data sets, modules, and passes. The false match weight distributions are skewed to either the right or left and the transformation does little to make the distributions more normal. The distributions for the true records are bi-modal and the transformation shifts the two spikes.

The graphs reveal there is little distribution in the weights, with huge spikes above the cut-off values. This suggests uniformity in data quality (in terms of, complete names, address, and dates of birth), showing that most were assigned the same composite weight.

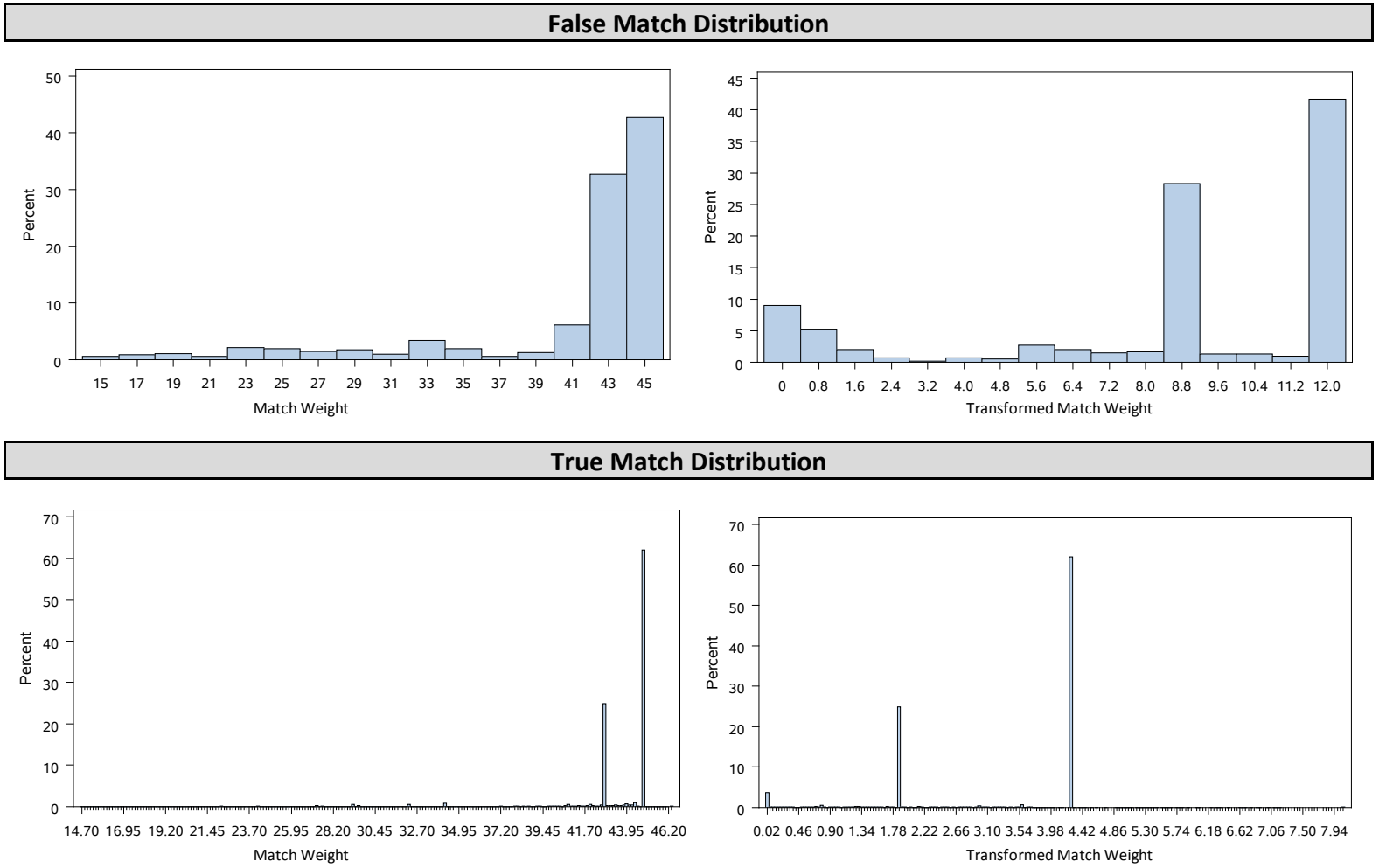
The transformed data are not normally distributed, but absent any other method in which to model false match rates, we proceeded with the Belin and Rubin method. Future research will consider other methods, which may prove to be more suited to the type of data we use in CARRA.

Figure 2. 2011 MEDB Before and After Box-Cox Transformation (Name Search, Passes 1-4)



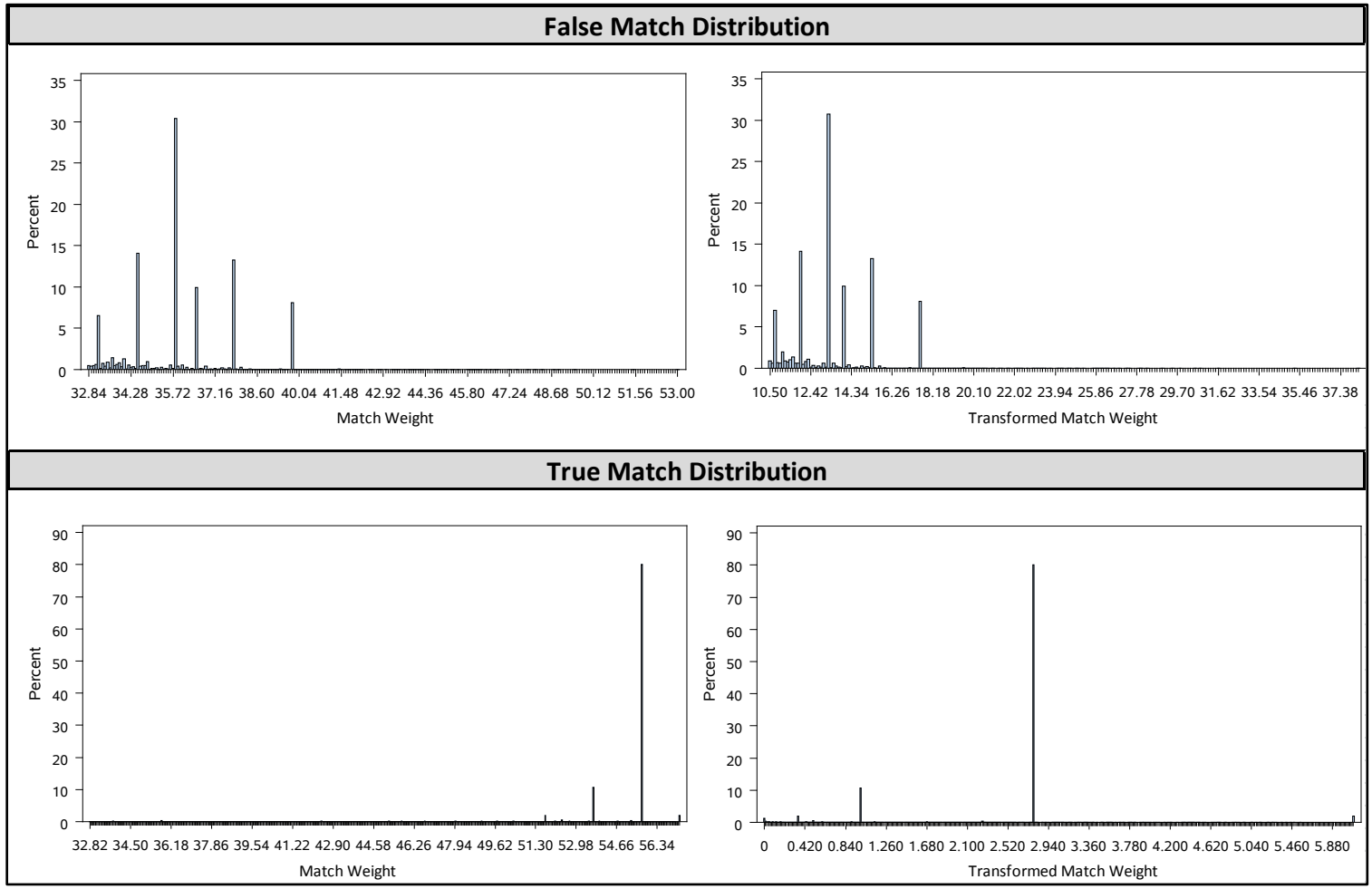
Source: 2011 MEDB.

Figure 3. 2011 IHS Before and After Box-Cox Transformation (GEO Search, Passes 1-4)



Source: 2011 IHS.

Figure 4. 2010 Vendor 2 Before and After Box-Cox Transformation (DOB Search, Passes 1-4)



Source: 2010 Vendor 2.

Table 5[11] presents estimated false match rates at cut-offs -- specified in the PVS search parameters -- using our application of the Belin and Rubin methodology, alongside observed percentages of false matches. The Belin and Rubin false match rate is a predicted probability of false match for two records as a function of the composite weight. Belin and Rubin stated their model produces conservative estimates and most likely over-estimates false match rates. Furthermore, we have determined this is not the best model for estimating false match rates in the PVS because of the non-normality exhibited in the distributions of true and false transformed weights.

The Belin and Rubin model estimated higher false match rates than observed for all Medicare groupings and 2010 commercial Geosearch. The Belin and Rubin model resulted in lower false match rates for the 2011 IHS GeoSearch, ZIP3 Search, Name Search and DOB Search, and 2010 commercial Name Search and DOB Search.

Table 5. Modeled and Observed False Match Rates at Cut-Off Weights

	2011 MEDB		2011 IHS		2010 Commercial	
	Probability of a False Match Rate at Cutoff	% Observed Error	Probability of a False Match Rate at Cutoff	% Observed Error	Probability of a False Match Rate at Cutoff	% Observed Error
GeoSearch[1]						
Passes 1-4; Cut-off=14.64	2.220%	0.004%	0.001%	0.046%	1.700%	0.146%
Zip3 Spatial Adjacency[1]						
Pass 2; Cut-off=32.13	3.550%	1.177%	0.261%	0.443%	NA[2]	1.446%
NameSearch						
Passes 1-4; Cut-off=32.14	2.230%	0.262%	0.272%	0.714%	2.550%	4.177%
DOBSearch						
Passes 1-4; Cut-off=32.83	4.050%	0.177%	0.106%	0.392%	1.670%	6.604%

Note:

[1] Passes 5-6 and 9 are omitted because of small sample sizes. Zip3 pass 1 omitted because of a low number of false matches.

[2] The model for this data, module, and pass did not converge and we are examining why.

Source: 2011 MEDB, 2011 IHS, 2010 commercial Vendor 1 and Vendor 2.

[11] We didn't include the first pass of ZIP3 because of the low number of false matches.

6. SUMMARY

We utilized two methods on six data sets (2011 MEDB, 2011 IHS, 2012 MEDB, 2012 IHS, 2010 Vendor 1, and 2010 Vendor 2) of estimating false match rates in the PVS. We were able to create truth data automatically and compared observed false match rates with modeled false match rates using Belin and Rubin's methodology. The observed match rates were lower in seven of the twelve test groupings. We have also shown that the Belin and Rubin method is not a good approach for our error research because of the non-normality of the true and false distribution of transformed weights.

We intend on testing the Belin and Rubin model further by using the PVS as it was run in production (records only proceed to the next module if they failed to find a PIK in a prior module), rather than processing each set of verified records through search modules independently. This can lead to higher error rates because a record that is assigned a false PIK in one module may have found the correct PIK in a previous module. Future research will test this hypothesis. We will also explore the impact of changing cut-off values that define matches from non-matches, comparing the false match rates before and after the adjustments.

This research used federal data with high quality person identifiers and commercial data that contained SSN data. This high quality person data facilitated matching; however, this level of quality isn't available for all files processed through the PVS. This made it possible to create truth data without involving clerical review. NORC (2013) recently completed PVS research which involved creating truth data for survey data. Currently, Census Bureau surveys do not collect SSN, so NORC used a method of creating truth different than this research. We will run our Belin and Rubin software on the NORC data to explore the method and resulting error rates.

More research is needed to estimate the false match rate for files processed through the PVS that don't contain any SSN data and have lower data quality. Future research will also examine different statistical methods to measure false match rates, including the use of Markov Chain Monte Carlo (MCMC) methods in the mixture model. These methods appear promising and may give better results than the Belin and Rubin model applied in this paper, but are computationally slow. Future work will compare MCMC methods to the Belin and Rubin estimates and observed error.

Appendix A – Blocking and Matching Variables

Each search module has its own parameter file, which includes number of passes (match attempts). For each pass the following parameters are defined: blocking variables, matching variables, matching weight for each match comparison and the cut-off value for the pass.

Blocking divides the comparison space into manageable pieces. For example, suppose File A contains 300 million records and File B (the reference file) contains 1.2 billion records. The comparison space is 300,000,000*1,200,000,000. This comparison is unmanageable because it would take too long and may not be possible with the available computer resources. When using a 5-digit ZIP Code as a blocking variable the data are cut into 5-digit ZIP Code chunks, and the search comparisons are done within these cuts. Typically, the blocking strategy starts more strictly and loosens restrictions to widen the search space with each subsequent pass.

Matching variables are the variables to be compared to determine whether the record from the input file and the record in the reference file are a match. The **matching weight** assigns a value, dependent upon the degree of agreement or disagreement between two matching variables. All of the comparison weights for the pass are summed to create a composite weight value, which is compared to the **cut-off value** for the pass. The cut-off value is the threshold number to which the composite weight is compared. If the composite weight is equal or greater to the cut-off value, the records are deemed a match.

The table (on the next page) describes the blocking and matching variables for each pass for all modules used for this research.

Blocking and Matching Variables For Search Modules

Module	Pass	Cut-off	Blocking Variables	Matching Variables
GeoSearch	1	14.64	MAFID	First 15 characters First Name First 15 characters Middle Name First 12 characters Last Name Generational Suffix Gender Day of Birth Month of Birth Year of Birth SSN (used in production – Not in this research)
	2	14.64	MAFID	**Switch First/Last Names** First 15 characters First Name First 15 characters Middle Name First 12 characters Last Name Generational Suffix Gender Day of Birth Month of Birth Year of Birth SSN (used in production – Not in this research)
	3	14.64	CARRA generated GEOKEY	Same as Pass 1 (GeoSearch)
	4	14.64	CARRA generated GEOKEY	Same as Pass 2 (Switch First/Last Names)
	5	13.89	House Number Soundex for Street Name	Same as Pass 1 (GeoSearch)
	6	13.89	House Number Soundex for Street Name	Same as Pass 2 (Switch First/Last Names)
	9	32.13	ZIP3 First 2 characters First Name First 2 characters Last Name Year of Birth	Same as Pass 1 (GeoSearch)

Module	Pass	Cut-off	Blocking Variables	Matching Variables
ZIP3 Adjacency Search	1		House Number (including apartment number) Soundex for Street Name	First 15 characters First Name First 15 characters Middle Name First 12 characters Last Name Generational Suffix Gender Day of Birth Month of Birth Year of Birth SSN (used in production – Not in this research)
	2		Day of Birth Month of Birth Year of Birth	Same as Pass 1 (ZIP3Adj Search)
Name Search	1	32.14	NYSIIS code for First Name NYSIIS code for Last Name Day of Birth Month of Birth Year of Birth	First 15 characters First Name First 15 characters Middle Name First 12 characters Last Name Generational Suffix Gender Day of Birth Month of Birth Year of Birth SSN (used in production – Not in this research)
	2	32.14	First character First Name First character Last Name Day of Birth Month of Birth Year of Birth	Same as Pass 1 (Name Search)
	3	32.14	First 2 characters First Name First 2 characters Last Name Year of Birth	Same as Pass 1 (Name Search)
	4	32.14	First 2 characters First Name First 2 characters Last Name Day of Birth Month of Birth	Same as Pass 1 (Name Search)

Module	Pass	Cut-off	Blocking Variables	Matching Variables
DOB Search	1	32.83	**Switch First/Last Names** First character First Name First character Last Name	**Switch First/Last Names** First 15 characters First Name First 15 characters Middle Name First 12 characters Last Name Generational Suffix Gender Day of Birth Month of Birth Year of Birth SSN (used in production – Not in this research)
	2	32.83	First 3 characters First Name First 3 characters Last Name	First 15 characters First Name First 15 characters Middle Name First 12 characters Last Name Generational Suffix Gender Day of Birth Month of Birth Year of Birth SSN (used in production – Not in this research)
	3	32.83	Reverse Soundex for First Name Reverse Soundex for Last Name	Same as Pass 2 (DOB Search)
	4	32.83	First 2 characters First Name Year of Birth	Same as Pass 2 (DOB Search)

References

Belin, T.R.and Rubin, D.B. (1995). “A Method For Calibrating False-Match Rates in Record Linkage,” Journal of the American Statistical Association, 430, pp. 694-708.

ESRI. December 6, 2012 from http://support.esri.com/en/knowledgebase/GISDictionary/search.

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

Fortini, M., Nuccitelli, A. and Liseo, B. (2002). "Modeling Issues in Record Linkage: A Baysian Perspective," Joint Statistical Meetings, August 2002.

Hall, R., Fienber, S. (1995). “Valid Statistical Inference on Automatically Matched Files,” Proceeding PSD'12 Proceedings of the 2012 international conference on Privacy in Statistical Databases Pages 131-142, Springer-Verlag Berlin, Heidelberg ©2012

Lahiri, P. A., and Larsen, M. D. (2005). “Regression Analysis with Linked Data,” Journal of the American Statistical Association, 100, 222-230.

Larsen, M. D., and Rubin, D. B. (2001), “Iterative Automated Record Linkage Using Mixture Models,” Journal of the American Statistical Association, 79, 32-41.

Larsen, M. D. (2005). “Hierarchical Bayesian Record Linkage Theory,” Iowa State University, Statistics Department Technical Report.

Larsen, M.D. (2010). Record Linkage Modeling in Federal Statistical Databases. FCSM Research Conference, Washington, DC.

Miller, C., Layne, M., Bouch, M., and Smith, D. (Upcoming Winter 2013). “ZIP3 Spatial Adjacency Search Module for the Person Identification Validation System (PVS),” Center for Administrative Records Research and Application (CARRA) Working Paper Series, U.S. Census Bureau.

McLachlan, Geoffrey, and Peel, David. *Finite Mixture Models.* John Wiley & Sons, Inc. (2000). Print. pp 19-21.

NORC at the University of Chicago (2013). "Task 4, Further PVS Research.”

O’Hara, A., Marshall, L. (2011). “2010 Census Evaluations, Experiments and Assessments Plan: 2010 Match Study,” 2010 Census Planning Memoranda Series, United States Census Bureau.

Scheuren, F., Winkler, W. (1993). “Regression Analysis of Data Files that Are Computer Matched – Part I,” Survey Methodology, 19, 1 pp. 39-58.

Tancredi, A., Liseo, B. (2011), “A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems.” Annals of Applied Statistics 5(2B), 1553-1585.

Wagner, D. (2012). "Documentation for the Multi-Match Record Linkage Software," CARRA Internal Working Paper, U.S. Census Bureau.

Wagner, D. and Layne, M. (2014). "The Person Identification Validation System (PVS),” Center for Administrative Records Research and Application (CARRA) Working Paper No. #2014-01, U.S. Census Bureau.

Winglee, M.,Valliant, R., Scheuren F, (2005). "A Case Study in Record Linkage," Survey Methodolgy, Vol. 31., No 1., pp. 3-11..

Winkler, W. E. (2007). “Automatically Estimating Record Linkage False Match Rates,” U.S. Bureau of the Census, Statistical Research Division Report, Research Report Series.