CARRA Working Paper Series

Working Paper #2014-01

The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software

Deborah Wagner
U.S. Census Bureau

Mary Layne
U.S. Census Bureau

Center for Administrative Records Research and Applications
U.S. Census Bureau
Washington, D.C. 2023

Report Issued: July 1, 2014

The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software

July 1, 2014

Deborah Wagner
U.S. Census Bureau

Mary Layne
U.S. Census Bureau

Abstract

The Census Bureau's Person Identification Validation System (PVS) assigns unique person identifiers to federal, commercial, census, and survey data to facilitate linkages across and within files. PVS uses probabilistic matching to assign a unique Census Bureau identifier for each person. The PVS matches incoming files to reference files created with data from the Social Security Administration (SSA) Numerical Identification file, and SSA data with addresses obtained from federal files. This paper describes the PVS methodology from editing input data to creating the final file.

TABLE OF CONTENTS

I. INTRODUCTION

The Census Bureau performs research with administrative records[1] files to investigate methods to improve the Census Bureau's statistical processes. Further, many projects at the Census involve matching persons across surveys and Federal data to enhance the understanding of participation in various Federal programs. Fundamental to this work is a method to ensure the same person is linked across multiple administrative files. The Census Bureau's Person Identification Validation System (PVS) is used to ascertain unique person and address identifiers.

The PVS uses probabilistic linking (Fellegi and Sunter, 1989) to match person data to a reference file. The reference files are derived from the Social Security Administration (SSA) Numerical Identification file (SSA Numident). The Numident contains all transactions recorded against one Social Security Number (SSN) 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.

The matched person record is assigned a unique person identifier called the protected identification key (PIK) and it 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.

The Census Bureau developed the PVS in 1999 in collaboration with the SSA. The PVS was tested using the 1997 Current Population Survey (CPS) survey previously verified using the SSA's SSN validation process, the Enumeration Verification System (EVS). An independent evaluation led by the Census Bureau's Demographic Surveys Division (DSD) compared the results of the PVS and EVS and showed the PVS achieved match rates more than two percentage points higher than achieved by SSA's EVS. The use of address data in the PVS system proved valuable in achieving this higher SSN assignment rate. As part of our Memorandum of Understanding with SSA, the Census Bureau must validate SSNs using PVS any time data are linked to any SSA administrative data. The SSA has authorized the use of the PVS at the Census Bureau.

---

[1] As used here, "administrative records" are files from the IRS, Department of Housing and Urban Developments Tenant Rental Assistance Certification System and Public and Indian Housing Information Center (HUD-TRACS and HUD-PIC), Selective Service System (SSS), Indian Health Service (IHS), Medicare Enrollment (MEDB), and commercial data from various sources.

The PVS was also used within the Statistical Administrative Records System (StARS) (Farber and Leggieri, 2002) to un-duplicate person records from Federal administrative records data. The StARS was produced annually from 1999 through 2010, each year processing approximately 900 million person records through the PVS system. PVS has been used with the Census Bureau's demographic surveys and censuses since 2001.

The PVS is the cornerstone of the Census Match Study (O'Hara and Marshall, 2011). Following the 2010 Decennial Census, the Census Match Study was undertaken to compare the coverage of administrative records and commercial data to the 2010 Decennial Census enumeration. In addition to data from Federal administrative records, commercial sources were evaluated. Nine sources of commercial data were selected to provide additional addresses and information for people not found in administrative records as of the Census date of 4/1/2010. These commercial data sources were thought to contain more timely information on address data as of census day to supplement addresses not available on the Federal files.

To meet the challenge of employing the PVS on commercial data sources and new Federal files, The Center for Administrative Records Research & Applications (CARRA) reviewed its current PVS capability. Independent contractors (NORC, 2011a) evaluated the PVS and recommended further enhancements to improve the already sound methodology of PVS.

One of the key enhancements increased the coverage of the reference files by including records for persons with Individual Taxpayer Identification Numbers assigned by the Internal Revenue Service (ITINs) to the SSN-based Numident data. The PVS is an important tool at the Census Bureau as it continues to pursue research using the 2010 Decennial data and plan for administrative records use in the 2020 Decennial Census.

Record linkage requires human input throughout the process. Files need to be properly edited before they can be linked successfully. Parameters have to be set and links examined. In the subsequent sections, we describe the PVS. Section II provides an overview of record linkage and Section III details the methodology for the current PVS system and its uses at the Census Bureau. Section IV presents PVS results for Federal and commercial files. Section V discusses advantages and future improvements of the PVS. Finally, Section VI summarizes the salient points of the paper.

II. RECORD LINKAGE BACKGROUND

Record linkage is the process of bringing together two or more records relating to the same entity. In 1946, H. L. Dunn of the United States National Bureau of Statistics introduced the term in this way: "Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Record linkage is the name of the process of assembling the pages of this Book into a volume" (Dunn, 1946). In 1959, computerized record linkage was first undertaken by the Canadian geneticist Howard Newcombe and his associates.

In 1969, Ivan Fellegi and Alan Sunter provided a mathematical framework to provide a viable theorem for linking two data files on common characteristics. This section briefly describes the general theory of record linkage. The paper by Felligi and Sunter (1969) provides the detailed theory.

Felligi and Sunter describe a *comparison space*, **A** x **B**, consisting of all comparisons of records from two files *a* and *b*. There are three outcomes for the comparison space **A** x **B**: links, possible links, and non-links. In order to classify record pairs, comparisons are made between the same fields in each of the files.

Denote the set of all comparison vectors, **Γ**, in **A** x **B** by:

$$\boldsymbol{\Gamma}[\boldsymbol{A}\ x\ \boldsymbol{B}] = \{\tau^1[(\alpha(a),\beta(b))],\{\tau^2[(\alpha(a),\beta(b))],\cdots,\{\tau^k[(\alpha(a),\beta(b))]\}$$

Two conditional probabilities are computed for each comparison pair.

m, is the probability of agreement for a given comparison, when the record is *in truth* a match. Because all matching variables are subject to data coding error (for example, typographical or scanning errors), this *m* probability is less then 1.0 .

u, is the probability of the comparison agreeing purely by chance for two records not belonging to the same individual.

The ratio of these two conditional probabilities, *R*, is an odds ratio defined as:

$$R = \frac{P(agree|\boldsymbol{m})}{P(agree|\boldsymbol{u})} = \frac{m(\tau)}{u(\tau)}$$

Probabilistic record linkage assigns a value of R, to a pair of records.[2] The analyst provides a lower and upper threshold for a match. The optimum linkage is the one

[2] In practice, the comparison space is far more complicated because matching occurs on multiple fields, e.g., first name, middle name, last name, street address, street name, date of birth, etc.

where records linked have a higher odds ratio than the upper threshold. All unlinked records have a odds ratio lower than the lower threshold. These two thresholds are set based on tolerance levels for the two types of error, linking unmatched records and failing to link matched ones.

If $R \geq Upper$, then the pair r, is a link. If $Lower < R < Upper$, the pair r is a potential match and assigned for analyst review. If $R \leq Lower$, the pair r is a non-match. The cut-off thresholds for *Lower* and *Upper* are determined before the linking is done. It is important to review records falling into each category.

Individual agreement weights, w_τ for the i^{th} fields of the r^{th} pair of records can be used as a substitute for the odds ratio, R:

$$w_\tau = log \frac{P(\tau \in \Gamma | \boldsymbol{m})}{P(\tau \in \Gamma | \boldsymbol{u})}$$

The analyst designates weight values for upper and lower cutoffs. Record pairs where weights are above upper cutoff are links, weights below the lower cutoff are non-links, and those weights between cutoffs are in the clerical review region.

Estimating *m* and *u* Probabilities

There are several methods for automatically or semi- automatically estimating the m and u probabilities (Winkler, 2002, Winkler and Yancey, 2006). Often, value specific or frequency-based estimates are employed. Another approach is using the EM algorithm, which involves finding the maximum likelihood estimate of parameters. The likelihood function is simplified by assuming the existence of, and values for, missing or hidden parameters (in this case, m and u) (Bilmes, 1998).

In practice, a single threshold value is often used (Winglee, 2005) and for PVS production purposes, CARRA sets a single threshold. In this case, the analyst decides links made below a threshold value are not valid links. Upper and lower bounds can also be set in PVS, and the system can provide links for analyst review.

III. PVS METHODOLOGY

The PVS software serves two primary functions: standard data editing and probabilistic matching. The PVS provides a documented, practical solution for processing many data files including census surveys (Current Population Study, American Community Survey), administrative records, and commercial files. The PVS probabilistically matches an incoming file to reference files in order to assign an anonymous PIK.

Optimized parameters, which provide information relevant to a probabilistic search,[3] are preset by analysts for each file type based on years of usage, research, and testing. The staff in CARRA have expertise in the parameter setting process from their research and bring this knowledge to production. When a new file arrives, different from all previous files, CARRA staff optimize the parameters for probabilistic linking. Record linkage is both an art and a science. It is a balancing act between link quality, processing speed, and setting optimal parameters.

3.1 Multi-Match Software

The PVS employs its probabilistic record linkage software, Multi-Match (Wagner 2012), as an integral part of the PVS. CARRA's Multi-Match is a generalized probabilistic matching routine at the core of all of CARRA's matching applications, and it is used heavily in the PVS. The Multi-Match is programmed as a SAS macro and is used by various record linkage applications. A user-defined parameter file provides information to Multi-Match, enabling it to perform the following functions

- Block records according to parameter file specifications
- Process records through passes,[4] as defined by the parameter file
- Perform comparisons between records in the incoming file and those in the reference file
- Create linked output.

Multi-Match can link *any* two files on *any* set of characteristics and can be used outside of the PVS construct for other matching applications.

Multi-Match, while an important part of the PVS, is not the only element in the PVS. The PVS performs many additional functions: data editing to prepare fields for matching, assigning Census geographical codes to incoming files, and data housekeeping functions.

3.2 Reference Files

Matching in the PVS requires a reference file to match against. Reference files are based on SSA's Numident and contain SSN, a CARRA assigned PIK associated with the SSN, date of birth, name, gender and any addresses where the person may have resided.

---

[3] Parameter files are discussed in-depth in Section 3.3.3.

[4] Each pass through the data defines differing matching strategies. See Section 3.3.3 for more details.

The SSA Numident file contains all transactions ever recorded against any single SSN. The Census Bureau builds a Census Numident on a regular basis from personal information derived from the SSA Numident file. All transactions related to a given SSN are resolved to produce a Census Numident file containing one data record for each SSN. All variants of name information for each SSN are retained in the Alternate Name Numident file, while all variants of date of birth data are retained in the Alternate DOB Numident. In addition to the Census Numident, PVS creates three other sets of reference files containing Numident data: the GeoBase Reference File, the Name Reference File, and the DOB Reference file.

The GeoBase Reference File appends addresses from administrative records attached to Numident data, including all possible combinations of alternate names and dates of birth for SSN. Addresses from administrative records are edited and processed through commercial software product to clean and standardize address data. ITIN data is also incorporated into the Geobase.

The Name and DOB Reference files are reformatted versions of the Census Numident and includes all possible combinations of alternate names and dates of birth, as well as ITIN data. All of the reference files contain SSN/ITIN and the corresponding PIK. When an input record is linked to a reference file, the corresponding PIK is assigned. Table 1 presents the number of observations in each of the reference files.

Table 1. Number of Observations in PVS Reference Files

Reference File	Observations
Census Numident	780 million
GeoBase	1.2+ billion
Name and Date of Birth	800 million

3.3 Preparing Incoming Files for PVS

The first step of the PVS process is to edit data fields to make them homogenous for comparisons between incoming and reference files. There are several standard edits, which insure maximum success during linkage.

3.3.1 Incoming File Edits

The first edits are parsing and standardizing - parsing separates fields into component parts, while standardizing guarantees key data elements are consistent (e.g., STREET, STR are both converted to ST). Name and address fields are parsed and

standardized as they are key linkage comparators. Figure 1 provides an example of how name and address are parsed into separate fields and then standardized.

During record linkage, a matching scheme might require that the first two letters of the first name must agree, the first four letters of the last name must agree, and the first two letters of the middle name agree, with similar rules for the address components. In Figures 1 and 2 the records are determined to be non-matches even though at first glance they may appear to be matches. This example demonstrates the importance of data parsing and standardization.

Figure 1. Data Parsing and Standardization In File Editing

Record #	Name	Address
1	Mr. Bob G. Smith, Jr.	2345 S. Main Street
2	Robert George Smith	2345 Main St.

Figure 2. Parsed and Standardized Through File Editing

Record #	Name					Address			
	Prefix	*First*	*Middle*	*Last*	*Suffix*	*House #*	*Pre-Directional*	*Street Name*	*Type*
1	Mr.	Robert	G	Smith	Junior	12345	South	Main	ST
2		Robert	George	Smith		12345		Main	ST

The data parsing and standardizing phase of PVS is extremely important for the matching algorithms to function properly. The PVS system incorporates a name standardizer (McGaughey, 1994), which is a C language subroutine called as a function within SAS. It performs name parsing and includes a nickname lookup table and outputs name variants (standardized variations of first and last names). For example, Bill becomes William, Chuck and Charlie becomes Charles, etc. The PVS keeps both the original name (Bill) and the converted name (William) for matching. PVS also has a fake name table to blank names such as “Queen of the House” or “Baby Girl.” The name data are parsed, checked for nicknames, and standardized.

The PVS editing process also incorporates an address parser and standardizer, written in the C language and called as a function within SAS (U.S. Census Bureau Geography Division, 1995). It performs parsing of address strings into individual output fields (see Figure 2), and standardizes the spelling of key components of the address such as street type. The PVS also incorporates use of a commercial product to update zip codes, and correct misspellings of address elements.

3.3.2 Assign Geographical Codes to Address

The PVS provides an additional address enhancement by matching records in the incoming file to Census Bureau's Master Address File (MAF) in order to assign a unique address identifier, the MAF Identifier (MAFID), and other Census geographical codes (e.g., Census tract and block). The MAFID is used in the PVS for search purposes and as a linkage key for administrative files. Then, addresses are matched to the Census Bureau's Topologically Integrated Geographic Encoding and Referencing Database (TIGER) to obtain Census geographical codes.

CARRA receives bi-annual deliveries of an extract of the Census Bureau's MAF. Prior to using the MAF in any search, the MAF records are processed through the same commercial product as used in file editing to update zip codes and correct any misspellings of address elements. The MAFMatch function in PVS matches input addresses to the MAF extract using the Multi-Match engine and attaches the MAFID and Census geographical codes (county, tract, and block). This process contains a series of blocking strategies:

1. Matching addresses using the full address, including the within structure unit number
2. Using rural route addresses
3. Using basic street address (BSA) without the within-structure unit number.

The full address pass normally blocks on house number and zip code, matching on house number, street name, street prefix and suffix type, directional, and within structure unit ID. The rural route address pass normally blocks on the zip code and box number, matching on the rural route ID. The BSA-level pass is similar to the full address pass, but ignores the within structure unit ID. Only geographical codes are retained from BSA level matches.

Next, addresses are matched to TIGER (Census Bureau, 2010), which is comprised of shapefiles[5], and includes information for the fifty states, the District of Columbia, Puerto Rico, and the Island areas (American Samoa, the Commonwealth of the Northern Mariana Islands, Guam, and the United States Virgin Islands). TIGER Census geographical codes are added to the address record. Matching to the TIGER is done through one five digit zip code blocking pass.

3.3.3 Parameter Files

PVS uses the same Multi-Match engine for each probabilistic search type. For each search module the analyst defines a parameter file, which is passed to Multi-Match. The parameter file includes threshold value(s) for the number of passes, blocking keys, and within each pass, the match variables, match comparison type, and matching weights. The pass number determines how many times a match will be attempted, with various configurations and rules for the matching variables. In general, each successive pass is less restrictive than the previous one.

Records must first match exactly on the blocking keys before any comparisons between the match variables are attempted. Each match variable is given an m and u probability, which is translated by MultiMatch as agreement and disagreement weights. The sum of all match variable comparison weights for a record pair is the composite weight. All record pairs with a composite weight greater than or equal to the threshold set in the parameter file are linked, and the records from the incoming file for these linked cases are excluded from all remaining passes. All Numident records are always available for linking in every pass. Any record missing data for any of the blocking fields for a pass skips that pass and moves to the next pass.

Because the PIK is intended to link the same person between files, the cutoff thresholds are set conservatively (i.e., higher than might normally be set in some record linkage applications).

3.3.4 Blocking Strategy

Potentially, each record in the incoming data can be compared against all records in the reference file. The number of comparisons grows quadratically with the number of records for matching (Baxter, 2003). Considering the size of incoming data processed for Census Bureau purposes (incoming files can have over 500 million records) this is computationally untenable. Therefore PVS incorporates blocking strategies, which are methods of reducing the search space. Similar records are grouped together using

[5] A popular geospatial vector data format for geographic information systems software.

information from record attributes to create a blocking key, which may be comprised of more than one data attribute. Analysts have to be thoughtful in the selection of blocking keys, because of the error characteristics of the attributes used. Fields prone to a high probability of error should not be used. The use of multiple passes can mitigate this concern if independent blocking fields are incorporated for each pass.

In the PVS, the blocking key for the first pass is highly restrictive and has a low probability of reporting error. (An example is an exact address match.) Subsequent blocking schemes can be less restrictive, but may produce weaker links. A feature of the PVS is its analysis program, which is run after any search program. This program provides a listing of links, by pass number, with the score and listing of the records matched from each file. This allows the analyst to examine links in a very compact and elucidating form. For example, if there are three passes, the analyst can review matches in the third (less restrictive) pass with the lowest score to determine if the parameter estimates are yielding sensible links.

The Multi-Match engine, when utilized for the PVS, is a one-to many matching system. The reference file is searched to find the best match for the input record. All reference file records are available for matching during each pass, but matched input records do not proceed to the next pass.[6]

3.3.5 Summary

The time required to determine the appropriate data cleaning edits should not be under estimated, as it can take days or even weeks. In 2011 CARRA received over two billion commercial records in nine different files. While preparing these data records for PVS was time-consuming, the editing programs are reusable for future vintages of the same file.

Once the comparison fields used in the matching modules have been cleaned, the file proceeds through the rest of the PVS modules. Running PVS modules from beginning to end (edit program to the very last program) is a fast process. The 2010 Census Unedited File (CUF), had 350 million records and processed through every PVS module, excluding MAF match and SSN verification, in 60 hours .

In the following sections, we discuss each module of the PVS application. The user can choose to employ each module for the incoming file, or exclude certain modules if data are not available. Research is ongoing about the impact of switching the order of

[6] The PVS system can also be run to produce multiple matches for each input record (matching with replacement).

the modules. Preliminary results indicate the current module ordering is optimal. As new modules are added, the research on module ordering will be repeated.

3.4 PVS Search Modules

The PVS consists of one exact match – SSN Verification - - and four search modules: GeoSearch, NameSearch, DOBSearch, and Household Composition Search (HHComp). Incoming records *cascade* through the PVS – and only records failing a particular matching module proceed on to the next module. If a record is assigned a PIK from the reference file in any of the modules, no further searches are conducted. Each module has its own set of user defined blocking factors and parameter thresholds.

3.4.1 Verification Module

If the input file has a SSN data field, it first goes through the 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 corresponding PIK is assigned. The SSN verification module is an exact match to SSN, so no parameter file is required for this step.

3.4.2 GeoSearch Module

The GeoSearch attempts to find SSNs or ITINs for incoming records that failed the verification module or without reported SSNs. This module links records from the incoming file to the GeoBase through blocking passes defined in the parameter file.

The typical GeoSearch blocking strategy starts with blocking records at the household level, then broadens the geography for each successive pass and ends at blocking by the first three digits of the zip code. The typical match variables are first, middle, and last names; generational suffix; date of birth; gender and various address fields.

The data for the GeoSearch module are split into 1,000 cuts based on the first three digits of the zip code (zip3) for record. The GeoSearch program works on one zip3 cut at a time, with shell scripts submitting multiple streams of cuts to the system. This allows for parallel processing and restart capability.

The GeoSearch module also incorporates the adjacency of neighboring areas with different zip3 values (Miller, Bouch, Layne, 2012). This can eliminate the bias of limiting

the blocking strategy to exact match on zip3 and obviates missing links based on zip3 blocking.

After the initial set of links is created, a post-processing program is run to determine which of the links are retained. A series of checks are performed. First the date of death information from Numident is checked. Next a check is made for more than one SSN assigned to a source record. If so, the best link is selected. If no best SSN is determined, all SSNs assigned are dropped and the next module begins. A similar post-processing program is run at the end of all search modules.

3.4.3 NameSearch Module

The NameSearch module searches the reference files for records failing the Verification and GeoSearch Modules. Only name and date of birth data are used in this search process. NameSearch 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 NameSearch blocking strategy starts with a strict first pass, blocking records by exact date of birth and parts of names. Successive passes block on parts of the name and date of birth fields to allow for some name and date of birth variation. The typical match variables are first, middle and last names, generational suffix, date of birth, and gender.

After the initial set of links is created, a post-processing SAS program is run to determine which of the links are retained, similar to the program used after the GeoSearch.

3.4.4 DOBSearch Module

The DOBSearch module searches the reference files for the records that fail the NameSearch, using name 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 DOBSearch 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.

3.4.5 HHCompSearch Module

The HHComp Search module searches the reference files for records that fail the DOBSearch. 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 persons with a PIK in the eligible household, all of the geokeys from the PVS GeoBase are extracted for each of these PIKs. The geokeys are unduplicated and all persons are selected from the PVS GeoBase with these geokeys. Next, the program removes all household members with a PIK, leaving the unPIKed persons in the household. This becomes the reference file to search against. There are typically two passes in this module. Records are blocked by MAFID, name, date of birth, and gender.

3.5 Master File Creation

Another feature of the PVS is the creation of a master file for the analyst. The master file creation is the last step in the PVS process. This program collects the results of each search module and creates a variable describing the final disposition of the PIK assignment – the module it was assigned in or the reason for the inability to assign a PIK (e.g., respondent's SSN is in a reference file, but the name is not a match). The program adds the PIK, MAFID and Census geographical codes to the incoming file and removes SSN or ITIN.

IV. PVS RESULTS

This section describes the results for some of the files CARRA has processed through the PVS system. The primary success measure for the PVS is the number of people matched to a reference file and assigned a PIK.

Table 2 presents detailed PVS results for three types of files: 2010 commercial data, a 2011 Federal representative file,[7] and the 2010 Census of Population and Housing (decennial census). In each of the three files, some records lack the necessary information to continue through the PVS. Ninety nine percent, 100 percent, and 97 percent of records for commercial, the Federal, and Census data have the necessary information to proceed through PVS.

The Decennial Census file doesn't contain SSNs, while the some of the commercial files and most of the Federal files do. The files with SSNs are eligible to go through the Verification module. In the commercial and the Federal file one record represents a person (person-level) and do not reflect household rosters. Therefore these files are not put through the Household Composition module. The Decennial Census persons are enumerated within households, and these records were submitted to the Household Composition module, provided they failed previous modules.

The SSNs in the Federal file are verified at a much higher rate than the commercial files. This seems reasonable, as participation in Federal programs requires accurate input. As show in Table 2, 99.9 percent of the Federal file records where verified leaving a scant .1 percent to proceed through the other probabilistic modules. In the commercial files, 76 percent of records were verified in the Verification module.

Because records for the 2010 Decennial Census lack SSNs, this file is sent to the probabilistic GeoSearch module and PIKS are assigned to 86 percent of the records. Of the 60,351 records sent to GeoSearch for the Federal file, 56 percent have a PIK assigned. The rate for the commercial files is 29 percent. The GeoSearch module assigned more PIKs to the census data than the commercial data because census addresses are generally cleaner and more complete.

[7] The commercial file results are the average of two commercial files and the Federal file is left unnamed.

Table 2. Sample PVS Results

PVS/MAFMatch Results	Average Of 2 Commercial Files	2011 Federal File	2010 Decennial Census
# Records delivered	339,295,722	53,181,072	312,471,327
# Addresses to MAFMatch	339,295,722	53,181,072	
Matched to MAF	283,605,634	47,721,226	
% with MAFID	83.59%	89.73%	
# Person Records	339,295,722	53,181,072	312,471,327
NO SEARCH: Refuse/OptOut (VERFLG=R,O)	0	0	0
NO SEARCH: Blank Name (VERFLG=X)	22,491	0	10,367,975
Persons Available for PVS	339,273,232	53,181,072	302,103,352
With SSN (any 9 digit number)	257,979,068	53,118,553	0
No SSN - to search	81,294,164	62,519	302,103,352
To VERIFICATION (with SSN)	257,979,068	53,118,553	0
Verified	195,571,749	53,058,202	0
NOT Verified - to search	62,407,320	60,351	0
Percent VERIFIED	75.81%	99.89%	0
To GEO Search	143,701,483	122,870	302,103,352
Found in GEO Search (VERFLG=S)	41,847,479	68,525	259,873,717
NOT found in GEO Search	101,854,004	54,345	42,229,635
Percent Found in GEO Search	29.12%	55.77%	86.02%
To NAME Search	101,854,004	54,345	42,229,635
Found in NAME Search (VERFLG=T)	5,328,014	13,158	19,960,452
NOT Found in NAME Search	96,525,990	41,187	22,269,183
Percent Found in NAME Search	5.23%	24.21%	47.27%
To DOB Search	96,525,990	41,187	22,269,183
Found in DOB Search (VERFLG=D)	73,802	587	304,690
NOT Found in DOB Search	96,452,188	40,600	21,964,493
Percent Found in DOB Search	0.08%	1.43%	1.37%
To HHComp Search	0	0	21,964,493
Found in HHComp Search (VERFLG=U)	0	0	1,975,295
NOT Found in HHComp Search	0	0	19,989,198
Percent Found in HHComp Search	0	0	8.99%
Total TO PVS (Avail for PVS)	339,273,232	53,181,072	302,103,352
Total validated	242,821,044	53,140,472	282,114,154
Total not validated (of avail)	96,452,188	40,600	19,989,198
% Validated (OF AVAILABLE)	71.57%	99.92%	93.38%
% Validated (OF ALL)	71.57%	99.92%	90.28%

The NameSearch and DOBSearch yield another 6 percent PIKs for the commercial data, 26 percent for Federal data, and 49 percent for the Census data. Only the Census file is sent through the household composition module, assigning PIKS for an additional 9 percent of the records .

In the last two lines of Table 2, we present the overall number of records that get PIKS. One line shows the percentage of the records available to be searched and the last line describes percentages for the entire dataset, including records not sent to any of the modules. The Federal file has the highest percent validated, the Census has the second highest. While commercial files have the lowest percent validated, the rate is above 70 percent. The reference files are largely SSN based, resulting in higher validation rates in Federal files.

The PVS provided the linking capability for the 2010 Census Match Study (O'Hara and Marshall, 2011). Without the ability to link persons across files via the PIK, the research for the Census Match Study could not be undertaken. Assigning MAFIDs to addresses in administrative data through the PVS process also proved invaluable to the study.

V. CONTINUOUS ENHANCEMENTS

The PVS is programmed in the SAS language, making it accessible to programmers and analysts alike. The few C language subroutines used are transparent to analysts because they are called within SAS programs. With the SAS expertise within the Census Bureau, code alteration and support requirements are met with ease.

The modular nature of the PVS is advantageous for setting parameter estimates. It is also beneficial if problems are encountered and the program needs to be restarted. Further the modular nature of the PVS allows the analyst to select modules depending on their source data. The PVS also supports parallel processing via scripting, which allows more than one process to run at a time (up to 80 on CARRA's existing Linux machine), greatly speeding up the process.

Each search module utilizes the same Multi-Match engine, eliminating the need to write specialized code for different linkage types (GeoSearch, NameSearch, etc.). The engine is written in SAS and called from each search module. Each of the modules passes a parameter file to the Multi-Match engine.

Finally, PVS offers analysts the ability to easily examine links in each module. Links are presented with the source record, the reference record, the matching fields, the pass in which the records were matched, and the score for the match. This provides the analyst with invaluable information for checking the results of the matching and to discern patterns in their data.

A criticism (NORC 2011b) of the production use of PVS (versus research usage) is that it sets a single threshold controlling the probability of false linkages, but no control for the probability that a link is in fact a mismatch. Setting one threshold means the analyst can only change the probability of false matches ***or*** the failure to link matched records, but not both. Testing and research have indicated that a single threshold performs extremely well.

CARRA is continually updating and improving the PVS. Research is under way to estimate overall false match and non-match rates in the PVS. Concomitant with this effort, CARRA is developing a method to estimate the probability of a link, in order to provide analysts with a measure of the certainty of a link. Other current and planned research projects include: implementing the EM algorithm to automatically generate m and u probabilities, and developing a model to determine false match rate and matching bias.

VI. SUMMARY

The PVS has been used for over a decade at the U.S. Census Bureau. To date, over 120 survey, administrative record, and commercial files have been processed in a production environment through the PVS. It is robust, proven, and flexible software that provides an end to end process.

The PVS is well maintained software, providing orderly steps for probabilistically assigning unique person and address identifers (PIKs and MAFIDs) to data. The PIK or MAFID facilitate linkage across files. The PVS is written in SAS, with some calls to C routines, and can be well understood by users. It makes use of generalized matching software, Multi-Match, and provides many tools for the analyst to check parameters and matching results.

Core expertise and institutional knowledge resides in CARRA for using and updating the software, setting parameters, and interpreting PVS results. There are opportunities to increase functionality of the current PVS. These challenges and issues are actively being researched by CARRA.

REFERENCES

Bilmes, J.A., (1998). “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.” Internation Computer Science Institute, U.C. Berkeley
http://ssli.ee.washington.edu/~bilmes/mypubs/bilmes1997-em.pdf

Colosi, R., (2004). “Evaluation 7:An Examination of Address Matching Software.” United States Bureau of the Census.

Dunn, Halbert L. (1946). "Record Linkage" (PDF). American Journal of Public Health 36 (12): pp. 1412-1416. doi:10.2105/AJPH.36.12.1412.

Farber, J., Leggieri, C. (2002). “Building and Validating the Prototypical Statistical Administrative Records System,” United States Census Bureau.

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

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2007). Data Quality and Record Linkage Techniques, New York, N. Y.: Springer.

Kelly, R. (1984). “Blocking Considerations for Record Linkage Under Conditions of Uncertainty,” United States Census Bureau, Statistical Research Division.

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. (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.

McGaughey, Anne (1994). “The 1995 Bureau of the Census Computer Name Standardizer.” Statistical Research Division.

Michelson, M. and Knoblock, C. A. (2006). “Learning Blocking Schemes for Record Linkage,” Proceedings of AAAI-2006.

Miller,C., Bouch, M., Layne M. (2012). “Spatial Adjacency Search Module for the PVS”.U.S. Census Bureau. Center for Administrative Records Research and Application (CARRA) Working Paper.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). "Automatic Linkage of Vital Records," *Science*, 130, 954-959.

NORC at the University of Chicago (2011a). "Assessment of the U.S. Census Bureau’s Person Idenfication Validation System.”

NORC at the University of Chicago (2011b). "Person Validation and Entity Resolution Conference Report.”

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.

Prevost, R., Leggieri, C. (1999). Expansion of Administrative Rececords Uses at the Census Bureau: A Long-Range Research Plan. *FCSM Research Conference*, Washington, DC.

Sadsinle, M., Hall, R., Fienberg, S. (2011). "Multiple Record Linkage: Generalizing the Felligi-Sunter Theory to More Than Two Datafiles," *JSM Proceedings.*

Wagner, D. (2012). "Documentation for the Person Identification Validation System," CARRA Internal Documentation, U.S. Census Bureau.

Wagner, D. (2012). "Documentation for the Multi-Match Record Linkage Software," CARRA Internal Documentation, 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. (1988). "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," Proceedings of the Section on Survey Research Methods, American Statistical Association, 667-671.

Winkler, W. E. (1990b). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," Proceedings of the Section on Survey Research Methods, American Statistical Association, 354-359.

Winkler, W. E. (1994). "Advanced Methods for Record Linkage," Proceedings of the Section on Survey Research Methods, American Statistical Association, 467-472.

Winkler, W. E. (1995). "Matching and Record Linkage," in B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, Colledge, M. A., and P. S. Kott (eds.) Business Survey Methods, New York: J. Wiley, 355-384 (also available at http://www.fcsm.gov/working-papers/wwinkler.pdf).

Winkler, W. E. (2003a), "Methods for Evaluating and Creating Data Quality," Proceedings of the ICDT Workshop on Cooperative Information Systems, Sienna, Italy, January 2003, longer version in Information Systems (2004), 29 (7), 531-550.

Winkler, W. E. (2006a). "Overview of Record Linkage and Current Research Directions," U.S. Bureau of the Census, Statistical Research Division Report, Statistical Report Series.

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

Yancey, W.E. (2007). "BigMatch: A Program for Extracting Probable Matches from Large Files," Statistical Division Research Report, http://www.census.gov/srd/papers/pdf/RRC2007-01.pdf.

Yancey, W.E. (2002), "Improving EM Parameter Estimates for Record Linkage Parameters," Proceedings of the Section on Survey Research Methods, American Statistical Association, CD-ROM (also report RRS 2004/01at http://www.census.gov/srd/www/byyear.html).

United States Bureau of the Census, Geography Division.(1995). "Address Standardizer Documentation."

United States Census Bureau. (2010). "2010 Census TIGER/Line Shapefiles."