This document was made publicly available by the U.S. Census Bureau in January 2024 via email at the request of FSRDC #2854 in support of Joung and Reeves (2025). As no other public reference or archival version exists, I host it here to support reproducibility and citation.

Recommended citation:

U.S. Census Bureau, Data Integration Division. 2007. *StARS Person Processing: Programming Specifications for Creation of the StARS Composite Person Record.* Washington, DC: U.S. Census Bureau. Available at: *https://andrewjoung.com/DataRepo/CPR_documentation.pdf*

Data Integration Division

StARS Person Processing Programming Specifications

Creation of the StARS Composite Person Record

September 6, 2025

Contents

Tables

1 Introduction

1.1 Project Overview

One of the final steps in creation of the most recent and final StARS database by the Data Integration Division ("DID") is the construction of a unique record for each person reflected in the seven national level administrative record files that comprise the StARS database. This unique record, entitled the Composite Person Record ("CPR"), is designed to reflect all known demographic characteristics and a unique address associated with a given person. Through record linkage and unduplication techniques, a single record is created and maintained in the StARS CPR File for persons with a verified Social Security Number (SSN).

Two separate processing streams are employed to develop the StARS database – one for addresses from the administrative records and another for persons (SSNs) reflected in the administrative records. The respective output files from each of the processing streams are brought together to create Linked Person Files that are used to create the CPR. The Linked Person Files contain the demographic and geographic variables required to build the CPR during, what is essentially, a final unduplication process. This specification identifies the programming actions required to create the Linked Person Files in preparation for building the CPR, and creation of the CPR where demographic and geographic selection rules are applied to develop a unique CPR.

1.2 Background

The most current and final StARS process relies on file linkage to obtain unique addresses and unique SSNs. A number of intermediate processing files, as well as production output files, are generated within the various StARS processing modules. The common thread to link the files is a set of record identifier variables that are defined in detail in previous StARS specifications. Definitions of the linkage variables are provided as follows:

1. **UID16** **Unique Record Identifier:** a 16-character record identification label is assigned to each record within the original input source files. Although address and person processing are split into two separate streams, the UID16 variable assigned during initial editing of the input source files serves as the linking variable between the two processes. An additional variable used in conjunction with the *UID16* is also assigned during initial file editing which identifies whether the record is processed through the address (*UID17A*) or person (*UID17P*) stream. The *UID16* variable is the key element for record sorting, unduplication, and linkage.
2. **AID** **Address Identification Number:** an 11-character identification label assigned to each address record within a given 3-digit ZIP Code file. The AID is comprised of the 3-digit ZIP Code followed by an 8-character sequence number.

3. **HUID** **Housing Unit Identifier:** a 35-character identification label assigned to each address residing on the Master Housing File after address record unduplication, address standardization, and other intermediate processing actions. The HUID also serves as a record identifier for sort, unduplication, and selection purposes as the 35-character field replaces the actual input street address found on the original source file.

4. **MAFID** **Master Address File Identifier:** A 9-character field (assigned by Geography Division during the geocoding operation) that represents a system generated number during assignment of the MAFID.

5. **TIGERID** **Topologically Integrated Geographic Encoding and Reference Identifier:** A 10-character field that identifies a specific geographic location within the TIGER digital representation database. TIGERID (based on the latest and most current StARS database) is based on a range of house numbers along a given street.

Each of the identifiers described above are used during unduplication, linkage, or best address selection procedures throughout the StARS process.

1.3 Summary of Revisions

Note that the Change History document which contains a list of changes in the 2008 version of the Composite Person Record Specification. The list of changes is provided to assist in the review and quality assurance effort against the production version of the 2008 processes.

2 Input

2.1 Valid Person Files

The Valid Person Files, which currently consists of twenty segments, and is created as a product of the SSN Search and Verification (S&V) processing module. The twenty segments[1], split by the last two digits of the SSN, are sorted by SSN to mirror the Census Numident and represent the person input to the Linked Person Files.

The file record layout is based on the Edited Person File ("EPF") record layout that was created during the initial file edit and input to the SSN S&V processing module. Prior to creation of the Linked Person File, the 20 segments of the Valid Person Files are split to original source file cuts to enable a merge process with the Linked Address Files.

2.2 Master Housing File and Master Pointer File

The Master Housing File ("MHF") and Master Pointer File ("MPF") are products of StARS Address Processing. The MPF is merged with the MHF by AID to create a temporary Linked Address File ("LAF") after obtaining an extract of UIDs, HUIDs, and certain geography variables (from the "MHF") before the 1,000 LAF files are split into original source file cuts.

2.3 Linked Person File

For each source file cut, the LAF are merged with the Valid Person Files to create Linked Person Files. Administrative records geographic and demographic selection rules are applied to records within the Linked Person Files to create unique composite person records for the population universe reflected in the StARS database. The final Linked Person Files are re-split into the Census Numident format of 20 segments (based on SSN) to facilitate PIK assignment (replacing the SSNs) and a merge with the Person Characteristics File. The record layout for the Linked Person Files is presented in Section 3.3.4 as input to the CPR building process.

2.4 Person Characteristics File (PCF)

The PCF, sorted by SSN, is segmented to match the Census Numident (and the Linked Person Files). The PCF contains race, gender, and mortality data for all person records contained in the Census Numident.

3 Processes

3.1 Overview

Prior to creation of the CPR, the final output from the address and person processing streams must be linked in order to create a complete listing (file) of addresses for each person record

[1] A 21st segment, containing all invalid or not found (unverified) records, and records with no SSN, are not eligible for inclusion in the CPR. Instead, the unverified SSN records are retained in their pre-collapsed form (sorted by UID) for review and research by other staff within the Data Integration Division.

(SSN) and its' associated PIK. The complete list of addresses for each SSN is known as the Linked Person Files. The CPR is created by selecting the "best" address and "best" demographic characteristics from the data available on the Linked Person Files. Once the selection criteria are processed from among the records available for each SSN on the Linked Person File, the "selected" data is written out to the CPR but organized by the PIK.

3.2 Process Outline – Program Definitions

The following programs are used to construct the CPR through the creation of several intermediate files and record linkage steps.

Programs 34 & 35 Create Linked Person File

The Final Valid Person Files, split by original source cut, are merged with the Linked Address Files to create Linked Person Files. The program is run twice: Program 34 for all source files *excluding* the 'J' records (IRS 1099 file) and Program 35 for all the 'J' source records. The records are output in a Numident-like 20 segments based on the last two characters of the SSN.

Program 36 Create Final Linked Person File

The Final Linked Person File results from an "append and sort" process of the linked files created in Program 34 and 35. The IRS 1099 Linked Person Files are appended to the "all other" source Linked Person Files by segment to produce the 20 segments of the Final Linked Person File. The records are appended and sorted by SSN and UID16 to facilitate creation of the CPR in later processes.

Program 37 Create the Composite Person Record

The Composite Person Record (CPR) is built from the Final Linked Person File. During this program, the address selection algorithm and demographic characteristic selection rules are applied against the array of records for a given SSN to select the "best" values for output to the CPR. The Person Characteristics File and the PIK Assignment File are also run within this program. An Initial Comprehensive Address File (ICAF) is also created and output for further processing. The ICAF reflects ***all*** addresses for a given PIK, including the SSN and this record view is retained for the CPR output.

Program 38 Create Linked Unverified Person File

A Linked Person File for the unverified SSNs is created and retained in a pre-collapsed form, i.e., no demographic or geographic selection rules are applied to pick a "best record" for an unverified SSN. The DPB may access the pre-collapsed file for evaluation and research purposes.

Note: The Linked Address Files are created in Programs 32 and 33 as the final process actions during creation of the most recent StARS Master Housing File and Master Pointer File. Details of the process flow may be found in the most recent StARS specification.

Programs and File Names

For all current and valid SAS program library names and aliases please see the corresponding Process Flows for this specification. These file references can be found by contacting the Data Preparation Branch within the DID organization.

3.3 Create Linked Person Files (Programs 34 – 36)

The Linked Person Files are created in preparation for building the CPR in Program 37. Due to the volume of records in the IRS 1099 source file, two programs are run: Program 34 for non-IRS 1099 records and Program 35 for the IRS 1099 records. The two files are then appended and sorted by *SSN* and *UID16* to create the Final Linked Person Files in Program 36. The Final Linked Person Files are retained as a permanent dataset for tally and evaluation purposes. The process flow for Programs 34 – 36 may be viewed on pages 34 -36 of the StARS Flowcharts

3.3.1 Program 34 – Create Linked Person File 1

The Final Valid Person Files (from Program 23) for each source are merged with the Linked Address Files (LAF) for the same source that were created in Program 33 as the final step in creation of the Master Housing File (MHF - see Master Housing Specification for processing details).

1. Merge the Final Valid Person Files for the non-J (IRS 1099) records with the Linked Address Files for the non-J records by *UID16* within each source cut for each of the StARS sources.
2. Output the files in 20 segments (as in the Census Numident) based on the last two characters of the SSN.

3.3.2 Program 35 – Create Linked Person File 2

Complete the same process for the IRS 1099 records ("J" records) using the following program and file names:

1. Merge the Final Valid Person Files for the non-J (IRS 1099) records with the Linked Address Files for the J records by *UID16* within each source cut for each of the StARS sources.
2. Output the files in 20 segments (as in the Census Numident) based on the last two characters of the SSN.

3.3.3 1. Program 36 – Create Final Linked Person File

1. Append the Linked Person File "J" records (**`lnkpers2`**) to the Linked Person File non-"J" records (**`lnkpers1`**) using the **`append link person`** program and sort the output file by *SSN* and *UID16* to create the Final Linked Person File in twenty segments.
2. Following the append/sort operation, output a counts file with the variables specified in Table.

Table 1. Append/Sort Counts File

Name	Type	Length	Description
fid	Char	1	File ID (original source file indicator)
mcut	Char	3	Original cut number within file ID
addrcnt	Numeric	8	# of unmerged address records
mrgcnt	Numeric	8	# of records merged (matched)
perscont	Numeric	8	# of unmerged person records
reccnt1-20	Numeric	8	# of output records for each segment
recout	Numeric	8	Total number of output records (sum of reccnt1-20)

3. The Final Linked Person File (LPF) contains the geographic data (address information) required to select the best address and build the CPR. The LPF record layout follows:

Table 2. Linked Person File Record Layout

#	Variable	Description	Type	Len
1.	UID16	Unique Record Identifier	Char	16
2.	UID17P	Output Person Number	Char	1
3.	SSN	Social Security Number (SSN)	Char	9
4.	SSNSRC	SSN Source (File Source) H = HUD PIC File I = IRS 1040 File J = IRS 1099 File M = Medicare File N = Census Numident File T = Indian Health Service File V = Selective Service File W = HUD TRACS File	Char	1
5.	ARFNM	Administrative Record (ADREC) First Name	Char	15
6.	ARFNMO	ADREC First Name Overflow Flag 0 = No Character Overflow 1 = Additional Characters in Name Field	Char	1
7.	ARMNM	ADREC Middle Name	Char	15
8.	ARMNM2	ADREC Second Middle Name	Char	15
9.	ARLNM	ADREC Last Name	Char	20
10.	ARLNMO	ADREC Last Name Overflow Flag 0 = No Character Overflow 1 = Additional Characters in Name Field	Char	1
11.	ARSUFFIX	ADREC Name Suffix Generation Flag Equivalent (from the name standardizer) Generation Flag Suffix (output value) 0 = Blank 1 = JR 2 = III 3 = IV 4 = SR	Char	3

Table 3. Linked Person File Record Layout (cont.)

#	Variable	Description	Type	Len
12.	STFNM	Standardized First Name	Char	15
13.	STMNM	Standardized Middle Name	Char	15
14.	STMNM2	Standardized Second Middle Name	Char	15
15.	STLNM	Standardized Last Name	Char	20
16.	STDECD	Standardized Deceased Flag (returned by standardizer) 0 = Not Deceased (default) 1 = Deceased	Char	1
17.	NAMESRC	Name Source H = HUD PIC File I = IRS 1040 File J = IRS 1099 File M = Medicare File N = Census Numident File T = Indian Health Service File V = Selective Service File W = HUD TRACS File	Char	1
18.	SEX	Gender Blank = No Data Present / not applicable 1 = Male 2 = Female	Char	1
19.	SEXSRC	Gender Source H = HUD PIC File I = IRS 1040 File J = IRS 1099 File M = Medicare File N = Census Numident File T = Indian Health Service File V = Selective Service File W = HUD TRACS File	Char	1
20.	CYDOB	Birth Date (Century and Year) Blank = No Data Present / not applicable CCYY format – valid range between 1886 – 2008 (inclusive)	Char	4
21.	MMDOB	Birth Date (Month) Blank = No Data Present / not applicable MM format – valid range between 01 – 12	Char	2
22.	DDDOB	Birth Date (Day) Blank = No Data Present / not applicable DD format – valid range between 01 -31	Char	2
23.	DOBSRC	Date of Birth Source H = HUD PIC File I = IRS 1040 File J = IRS 1099 File M = Medicare File N = Census Numident File T = Indian Health Service File V = Selective Service File W = HUD TRACS File	Char	1

Table 3. Linked Person File Record Layout (cont.)

#	Variable	Description	Type	Len
24.	DOBSCOR	Date of Birth Score (* = non blank // - = blank) YYYY MM DD Score * * * 14 * * - 12 * - * 10 * - - 08 - * * 06 - * - 04 - - * 02 - - - 00	Char	2
25.	PRIM	Record of Primary Person Flag Value Bit Meaning Set 0 none NA (default setting) 1 0 Primary Person(s) 2 1 Secondary Person(s) 3 0,1 Primary/Secondary Persons Combination 4 2 Dependent Person(s) 5 0,2 Primary/Dependent Persons Combination 6 1,2 Secondary/Dependent Persons Combination 7 0,1,2 Primary/Secondary Dependent Persons Combination	Char	1

26.	HISP	Hispanic Origin // Blank = No Data Present / not applicable
		1 = Hispanic
		2 = Not Hispanic
			Char	1

27.	HISPSRC	Hispanic Source H = HUD PIC File I = IRS 1040 File J = IRS 1099 File M = Medicare File N = Census Numident File T = Indian Health Service File V = Selective Service File W = HUD TRACS File	Char	1

28.	RACE	Race // Blank = No Data Present / not applicable
		1 = White
		2 = Black
		3 = American Indian, Eskimo, or Aleut
		4 = Asian or Pacific Islander
		5 = Other
		6 = Unknown
			Char	1

29.	RACESRC	Race Source H = HUD PIC File I = IRS 1040 File J = IRS 1099 File M = Medicare File N = Census Numident File T = Indian Health Service File V = Selective Service File W = HUD TRACS File	Char	1

Table 3. Linked Person File Record Layout (cont.)

#	Variable	Description	Type	Len
30.	DECSSRC	Deceased Source H = HUD PIC File I = IRS 1040 File J = IRS 1099 File M = Medicare File N = Census Numident File T = Indian Health Service File V = Selective Service File W = HUD TRACS File	Char	1
31.	DCSTYP	Deceased Indicator Type - = blank (no data present) * = indication person is deceased from the following sources: AN = ADREC Name Standardization AV = ADREC Deceased Variable PD = PCF Actual Date of Death NV = Numident Deceased Variable AN AV PD NV Bits Set Value * * * * 0,1,2,3 15 * * * - 1,2,3 14 * * - * 0,2,3 13 * * - - 2,3 12 * - * * 0,1,3 11 * - * - 1,3 10 * - - * 0,3 09 * - - - 3 08 - * * * 0,1,2 07 - * * - 1,2 06 - * - * 0,2 05 - * - - 2 04 - - * * 0,1 03 - - * - 1 02 - - - * 0 01 - - - - 00	Char	2
32.	CYDOD	Date of Death (Century and Year) Blank = No Data Present / not applicable CCYY format – valid range between 1886 – 2008 (inclusive)	Char	4
33.	MMDOD	Date of Death (Month) Blank = No Data Present / not applicable MM format – valid range between 01 –12	Char	2
34.	DDDOD	Date of Death (Day) Blank = No Data Present / not applicable DD format – valid range between 01 -31	Char	2
35.	CYADCYDA	ADREC Cycle Date (Century and Year) CCYY format – valid range between 1886 – 2008 (inclusive)	Char	4
36.	MMADCYDA	ADREC Cycle Date (Month) Blank = No Data Present / not applicable MM format – valid range between 01 –12	Char	2
37.	DDADCYDA	ADREC Cycle Date (Day) Blank = No Data Present / not applicable DD format – valid range between 01 -31	Char	2

Table 3. Linked Person File Record Layout (cont.)

#	Variable	Description	Type	Len
38.	NMSCORE	Name Score (Calculated from ADREC name) (Key: **F** = Full **I** = Initial ***** = non blank **-** = blank) Last First Middle SFX Score - - - - 00 - - - * 01 - I - - 08 - I - * 09 - I I - 10 - I I * 11 - I F - 14 - I F * 15 - F - - 24 - F - * 25 - F I - 26 - F I * 27 - F F - 30 - F F * 31 * - - - 32 * - - * 33 * I - - 40 * I - * 41 * I I - 42 * I I * 43 * I F - 46 * I F * 47 * F - - 56 * F - * 57 * F I - 58 * F I * 59 * F F - 62 * F F * 63 Indicated Fields will yield a calculated score as follows: Suffix = + 1 Middle Initial = + 2 Middle Full = + 6 First Initial = + 8 First Full = + 24 Last Name = + 32	Char	2
39.	VERFLG	SSN Verification Flag // Set to **0** (zero - default value) for the Edited Person File	Char	1
		During the SSN S&V process, the following values are assigned **0** **Not Verified** Default value (zero) retained for records not eligible for Verification **A** **Not Verified** EPF record's SSN does not exist in the Census Numident **B** **Not Verified** EPF record's SSN exists in the Census Numident, but name does not verify **I** **Verified** IRS 1040 EPF record verified using special 1040 procedures **J** **Verified** IRS 1099 EPF record verified using special 1099 procedures **M** **Not Verified** EPF record not found because multiple SSNs found for a single person record **P** **Verified** SSN on EPF record verified using IRS 1099 child parent procedures **S** **Verified** EPF record received an SSN in the GeoKey search **T** **Verified** EPF record received an SSN in the name search **V** **Verified** EPF record is verified		
40.	AID	Address Identifier	Char	11
41.	MAFID	Master Address File Identification Number	Char	9
42.	HUID	Housing Unit Identifier	Char	35
43.	CONFDFGM	Confidence flag – MAF matching	Char	1
44.	CONFDFGT	Confidence Flag – TIGER Matching	Char	1
45.	FIPST2K	Census 2000 Tabulation FIPS State Code	Char	2
46.	FIPCTY2K	Census 2000 Tabulation FIPS County Code	Char	3

Table 3. Linked Person File Record Layout (cont.)

#	Variable	Description	Type	Len
47.	TRACT2K	Census 2000 Tabulation Tract	Char	6
48.	BLOCK2K	Census 2000 Tabulation Block (includes block suffix)	Char	5
49.	FIPST	Legal Geography Current FIPS State Code	Char	2
50.	FPCTY	Legal Geography Current FIPS County Code	Char	3
51.	MCD	Minor Civil Division - 2008	Char	3
52.	PLACE	Current Place Code	Char	4
53.	AIANA	American Indian / Alaska Native Code	Char	4
54.	CTST	Current Tabulation State	Char	2
55.	CTCTY	Current Tabulation County	Char	3
56.	TRACT	Current Tabulation Tract – 2008	Char	6
57.	BLOCK	Current Tabulation Block – 2008	Char	4
58.	BLKSFX	Current Tabulation Block Suffix – 2008	Char	1
59.	ZIP53	First three digits of 5-digit ZIP Code	Char	3
60.	ZIP52	Last two digits of 5-digit ZIP Code	Char	2
61.	ZIP4	ZIP + 4 Code	Char	4
62.	GEOSOFLG	Geocoding source flag	Char	1
63.	CENADDR	Good Census Address Indicator	Char	1
64.	ADDRTYPE	Address Type	Char	1
65.	GEOTYPE	Geocoding Type	Char	1
66.	BSACOD	Commercial Flag Indicator	Char	1
67.	BSASRC	Commercial Flag Basic Street Address Source Indicator	Char	3
68.	ATHOME	Work At Home Flag	Char	1
69.	PROXYFLG	Proxy Address Indicator Flag	Char	1
70.	USPSCD	USPS Indicator Code (from Code-1 processing)	Char	1
71.	MCDFP	FIPS MCD/CCD Code	Char	5
72.	PLACEFP	FIPS Place Code	Char	5
73.	AIANAFP	FIPS American Indian/Alaskan Native Code	Char	5

3.4 Program 37 – Creation of the Composite Person Record

The Linked Person Files stand ready for application of the demographic and geographic selection rules to create the Composite Person Record. Creation of the CPR is the final step in the StARS Person Processing module. Demographic and geographic variables must be selected from the Census Numident Files and output to the CPR. The Linked Person Files is read in and sorted in SSN order. The sort order will allow for an "SSN or PIK by group" processing methodology based on determining whether the current SSN read-in for processing differs from the previous SSN. The Linked Person File and the PCF will be processed in sequence beginning with segment 01.

The program to create the final CPR file generates the following four output files (the process flow may be viewed on page 37 of the StARS flowchart:

- CPR in 20 segments
- Crosswalk File in 20 segments
- ICAF in 20 segments
- Counts File

3.4.1 CPR Initial Processing Steps

Read in the first record and set an array for each of the demographic variables and the address (HUID) for retention and comparison of all records (by PIK) to determine the best variable value for output to the CPR. Variable processing is by PIK group. Retain all records and establish temporary processing variables to determine the variable tallies, source conflict and agreement, identification of the variable selection decision rule, and output the selected variable value. The processing steps are defined as follows:

1. Establish a processing array that eventually will establish a set of final variables:
    - Race
    - Hispanic origin
    - Gender
    - Date of Death (DOD)
    - Date of Birth (DOB)
    - Selected name fields
    - Address (HUID)
2. Obtain the PCF values identified in Table 3 for each PIK record read-in for processing. The Protected Identification Key (PIK) variable and *BESTRACE* variables are moved to the CPR output record. All other variables are used in intermediate processing steps to determine the CPR output value, when required. Using the PCF ensures a race, gender, Hispanic origin, or deceased value will be output for **each** record reflected in the CPR.

Table 3. PCF Variables Used in CPR Processing

Field	Type	Len	Description
SSN	Char	9	Social Security Number
PIK	Char	9	Protected Identification Key (value moved to the CPR)
GENDER	Char	1	Sex
BESTRACE	Char	1	Best Last Race (value moved to the CPR)
CH	Char	1	Census Hispanic Origin, recoded to Y = Hispanic or N = Non-Hispanic
IR	Char	1	Census Race
MH	Char	1	Modeled Hispanic Origin
MR	Char	1	Modeled Race
DODYYO	Char	4	Date of Death - Century & Year (from Numident, Death Master File or other source)
DODMMO	Char	2	Date of Death - Month
DODDDO	Char	2	Date of Death - Day

3. Determination of Race and Hispanic origin is primarily based on most frequent occurrence. Due to reporting inconsistencies among and within administrative records regarding race and Hispanic origin, three race and two Hispanic origin values are output to the CPR – an administrative record value (*ARRACE* and *ARHISP*), a PCF value (*LKLY_RACE* and *LKLY_HISP*) and a Census Numident race value (*BESTRACE*). The *ARRACE* / *ARHISP* values are only output to the CPR when there is a clear indicator of a reported value (as defined for each variable). A PCF value for race and Hispanic origin is output to the CPR for every record on the PCF.
4. Initialize the following output variables on the CPR. Zero fill the file source variable and blank fill all remaining variables upon initialization.
    a. File Source (*ARSOURCE*) – zero fill.
    b. Race (*ARRACE* and *LKLY_RACE*) – blank fill.
    c. Hispanic Origin (*ARHISP* and *LKLY_HISP*) – blank fill.
    d. Gender (*SEX*) – blank fill.
    e. Date of Death (*DOD*) – blank fill.
    f. Date of Birth (*DOB*) – blank fill.
    g. Name (selected name fields – see record layout) – blank fill.
    h. Address (*HUID*) – blank fill.
    i. Primary Person Record Flag (*PRIM*) – blank fill.
5. Check the input Verification Flag (*VERFLG*) to set the Verification Type (*VERTYPE*) value output to the CPR. The *VERFLG* value indicates, to some degree, the method of SSN verification. Compute the *VERTYPE* value in a three step process as follows:

a. For each input record, create a set of six 0 – 1 bit variables, numbered in the positions shown in Table 4 For each possible input *VERFLG* value, one bit variable is set to a value of **1** and all remaining bit positions set to **0**.

b. Combine the values (tally) across the records such that if a bit variable is ever a value of **1** for any of the input records, the result is **1** for that variable. If a bit was never a value of **1** for any of the input records, the result is **0**.

c. Concatenate the six resulting bit variables into a six-digit binary number where bit 1 is the least significant digit (2^0) and bit 6 is the most significant digit (2^5). Convert the binary number into a zero filled decimal (integer) and output the converted value to the *VERTYPE* field in the CPR.

d. As an example:

1) If the first of three records within a given SSN array reflect a *VERFLG* value of **J**, bit position 2 is turned on resulting in a bit set of **000100**.

2) If the *VERFLG* value for the second record is also **J**, the bit set is not changed since bit position 2 is already on.

3) If the *VERFLG* value for the third record in the SSN array is **T**, bit position 5 is turned on to yield a final bit set of **100100** for this SSN array.

4) Converting the binary number to an integer value for output to the CPR results in a value of **36** in this example.

Table 4. VERTYPE Bit Set Values

If *VERFLG* Input Value =	Turn on bit in position:	Definition
V	0	SSN found in Verification Process
I	1	SSN Verified using special IRS 1040 procedures
J	2	SSN Verified using special IRS 1099 procedures
P	3	SSN verified using IRS 1099 child/parent procedures
S	4	SSN found in the GeoKey Search Process
T	5	SSN found in the Name/DOB Search Process

6. The variable decision flag (***xxx****DECFL* – where ***xxx*** equals the output demographic variable) is a 1-character field that reflects the actual selection rule used to output the CPR demographic variable value. For each selection rule, a specific decision flag value is prescribed in the text accompanying the selection rule specification. The selection rules and flag values are also displayed in the record layout.

3.4.2 Variable Selection Rules

Once the last SSN of a group is read-in, processing arrays set, output variables initialized, and counters incremented, begin selection of the best value for the required variables within each SSN group using the selection rules specified for the variables described as follows:

A. Race

The basic criterion for administrative record race selection is mode. The race that is seen most frequently among all the SSN records under consideration for output to the CPR is generally the race assigned. One notable exception is where an Indian Health Service record reflects a race of American Indian (AI) or Alaska Native (AN). In this case, the AI/AN race is assumed to be the most likely race and no further check of other administrative records is required to determine the race value that is output to the CPR.

1. Determine the total records for each race, noting that Numident race values are ***not*** considered when determining the output *ARRACE* value.
2. Select the *ARRACE* value for output to the CPR based on the following rules:
   a. **Rule 1:** If any IHS source record reflects a race of American Indian (AI) or Alaska Native (AN), set the selected race to AI/AN on the CPR (value 3), and set the race decision flag (*RACDECFL*) value to 1.
   b. **Rule 2:** Select the most frequent race among the records, and set the *RACDECFL* value to 2. Input values of "blank, unknown, or other" are not considered under Rule 2.
   c. **Rule 3:** If ties occur among the most frequent observation ***or*** the only input race value available is "blank, other, or unknown", output a blank value for the *ARRACE*, and set the *RACDECFL* value to 3.
3. Select the *LKLY_RACE* value for output to the CPR based on the following rules:
   a. **Rule 1:** Set the *LKLY_RACE* value equal to the *IR* variable in the PCF, when available. The *IR* variable represents the Census race response value. Set the Census race decision flag (*CERADEFL*) value to **1**.
   b. **Rule 2:** If the *IR* value is blank, set the *LKLY_RACE* value equal to the PCF modeled race value (variable *MR*), and set the *CERADEFL* value to **2**.
4. Loop back through the input race data array to set the agree/disagree values for the *RACESRC* field in the CPR. The source conflict variable is an 8-character field where each position represents the original source file of the record. The value output to each field designates the *ARRACE* agreement/disagreement among records within a given source file. Character positions equate to the following source input records:

   (1) HUD PIC records (source file designator **H**)
   (2) IRS 1040 records (source file designator **I**)
   (3) IRS 1099 records (source file designator **J**)
   (4) MEDB records (source file designator **M**)
   (5) Census Numident records (source file designator **N**)
   (6) IHS records (source file designator **T**)
   (7) SSS records (source file designator **V**)
   (8) HUD TRACS records (source file designator **W**)

5. The possible values and definitions within each position are as follows:

Table 5. Variable Source Conflict Values

Value	Definition
0	No value for this variable for this SSN within source
1	All records agree (within source file)
2	All records disagree (non-blank) within source file
3	Conflict within source file (2 + 1)
4	All Input records are blank
5	Within source agree with selected value and blank (4 + 1)
6	Within source disagree with selected value and blank (4 + 2)
7	With in source agree, disagree, & blank (4+2+1)

B. Hispanic Origin

The Hispanic origin output to the CPR includes two variables – the administrative record value (*ARHISP*) and the PCF value (*LKLY_HISP*). As with race, Hispanic origin is also based on mode, noting that the Numident value is **not** considered in the selection process.

1. Select the administrative record Hispanic origin (*ARHISP*) value as follows:
   a. **Rule 1:** Select the administrative record Hispanic origin output (*ARHISP*) based on the most frequent non-blank observation and set the *HISDECFL* value to **1**.

   b. **Rule 2:** If ties occur among the most frequent observation ***or*** the only input race value available is "blank, other, or unknown", output a blank value for the *ARHISP*, and set the *HISDECFL* value to **2**.

2. Select the PCF Hispanic origin (*LKLY_HISP*) value as follows:
   a. **Rule 1:** Set the CPR *LKLY_HISP* value equal to the Census Hispanic origin response value (variable *CH*) and set the *CHSPDEFL* value to **1**.

   b. **Rule 2:** If the *CH* variable value is blank, set the CPR *LKLY_HISP* value equal to the PCF modeled Hispanic origin value (*MH*) variable, and set the *CHSPDEFL* value to **2**.

3. Loop back through the Hispanic data array to determine the source agree/disagree values (*HISPSRC*) as prescribed for *ARRACE* in the previous section.

C. Gender

1. Determine the total number of records that reflect a reported gender value.

Rule 1: If any Selective Service record is present, set the SEX variable value to 1 (male), and set the SEXDECFL value to 1.

   b. **Rule 2:** If no Selective Service record is present, select the non-blank gender value based on the most frequent observation and set the *SEXDECFL* value to **2**. Count the Numident value only once.

c. **Rule 3:** If ties occur among the observations or the input value is blank, move the PCF gender value to the output file and set the *SEXDECFL* value to **3**.

Note: The PCF gender values must be recoded to the CPR *SEX* value as follows:

If PCF Gender =:	Set CPR SEX Value to:
Blank	Blank
0 (zero)	Blank
M	1
F	2

2. Loop back through the gender data array to determine the source agree/disagree values (*SEXSRC*) as prescribed for *RACE* in a previous section.

D. Date of Death (DOD)

1. DOD from a Medicare record always consists of century, year, month, and day values (variable names -- *CYDOD*, *MMDOD*, and *DDDOD*). Incomplete dates of death will only appear on a Numident record.
2. Select the DOD based on the following rules:
   a. **Rule 1:** If only one Medicare record reflects a DOD, output the DOD to the CPR and set the *DODDECFL* value to **1**.
   b. **Rule 2:** If more than one Medicare record reflects a DOD, select the DOD from the Medicare record with the most recent cycle date and set the *DODDECFL* value to **2**.
   c. **Rule 3:** If no Medicare record exists, select the Numident DOD (if available), and set the *DODDECFL* value to **3**.
   d. **Rule 4:** If no DOD is available (all blank input values), check the PCF date of death fields (*DODYYO* [century and year], *DODMMO* [month], *DODDDO* [day]) for non-blank values. If present, move the date fields to the corresponding DOD fields in the CPR and set the *DODDECFL* value to **4** to indicate date of death information is present on the PCF. If no DOD is available on the PCF, set the *DODDECFL* value = blank to indicate no data present.
3. Loop back through the DOD data array to determine the source agree/disagree values (*DODSRC*) as prescribed for *RACE* in a previous section. Compare the entire DOD date fields (*ccyymmdd*) to set the agree/disagree value.

E. Date of Birth (DOB)

The selected DOB value for output to the CPR is based on the reliability of the source file. Table 6 displays the source file priority from the most to least reliable file source.

The Numident DOB value (if any) is used to fill all IRS 1040 and IRS 1099 record DOB fields, as well as any other source record where the input DOB field was blank or missing.

1. Select the output value to the CPR based on the following selection rules:

a. **Rule 1:** Select the output DOB based on the highest DOB score within the source file priority displayed in Table 6. Set the *DOBDECFL* value based on the source file priority value as indicated. If the highest DOB score is present on multiple records within the selected source, output the first record read-in among the ties.

Table 6. DOB Source File Priority

If Record Source File indicator is:		Set DOBDECFL Value to	If Record Source File indicator is:		Set DOBDECFL Value to
M	Medicare	1	**W**	HUD TRACS	4
V	Selective Service	2	**H**	HUD PIC	5
N	Census Numident	3	**T**	Indian Health Service	6

b. Rule 2: If the input record reflects no DOB, output blank values to the DOB fields in the CPR and set the DOBDECFL value to 7.

2. Loop back through the DOB data array to determine the source agree/disagree values (*DOBSRC*) as prescribed for *RACE* in a previous section. Compare only the century and year DOB date fields (*ccyy*) to set the agree/disagree value.

F. Name Fields

The basic premise for selecting name fields for the CPR is retention of the most complete **and** most recent name. The name score and administrative record cycle date present in the Linked Person File are used to determine the CPR name fields. Note however, the name fields are removed from the CPR for policy and privacy reasons. The name fields are retained on the Crosswalk File, along with the PIK and SSN, and housed in a restricted access directory. Process the CPR name selection in the following manner:

1. Determine the total number of records within the name array.
2. Sort the name array in descending order by cycle date and name score. The sort order ensures the first record among the sorted records reflects the most recent and the most complete last name.
3. Select the first record in the sorted array as the best name.
4. Loop back through the array to select the most complete name matching the last name of the best name selected in Step 3 (if any). Once, the best name is determined, move **all** name fields (both administrative record [*AR*] names and standardized names) from the selected Linked Person File record to the CPR name fields.
5. If another last name within the array shares the same cycle date as the selected name, set the name decision flag (*NMDECFL*) to **2**. Otherwise, set the *NMDECFL* to **1**.

Note: The name decision flag is set to indicate differing names in the name array, rather than indicating a particular selection rule used to select the "best" name retained on the CPR.

6. Loop back through the name array to determine the source agree/disagree values (*NAMESRC*). Set the values for each source file based on a match of the last name field (only) using a string comparator. Set the value to "agree", if the string comparator value ∃ 0.85. Otherwise, set to "disagree".

G. Address Selection

A logistic regression model, developed for the 2002 version of StARS, is used to select the best address to enhance the probability of selecting a "correct *and* best" address based on the model properties. Previous versions of StARS selected a best address by stepping through a series of prioritized criteria that primarily deferred to selecting a geocoded address over a non-geocoded address. The logistic regression model (LRM) used many of the same criteria as indicator variables to determine address validity. The indicator variables were factored into the LRM algorithm to establish a set of parameters that results in a scoring system for each address. Variables or data from the Linked Person Files used in the algorithm include the following:

HUID BSACOD ATHOME PROXY UID16

A Substructure variable is created from characters 24–35 of the HUID

A cycle date variable is created from 3 of the 7 source files

In a logistic regression model, the estimated probability is determined by the equation:

$$(1) \qquad probability\ of\ address\ invalidity = \frac{1}{1+e^{-l.e.}}$$

where

$$(2) \qquad l.e.\,(linear\ estimator) = \sum_{i=1}^{n} parameter_i \cdot estimator_i$$

and n = the number of estimators

Note that in the above equation, as the linear estimator becomes larger, the estimated probability also becomes larger. Thus, when comparing two linear estimators, the larger always generates a larger probability than a smaller. The property of this relationship is often described as one of a monotonically increasing function. Because of the monotonically increasing property of the logistic regression function, the linear estimator can be considered a score for an address. Where the score is low at a particular level of geography, the address is more likely to be correct at that level. Thus, when comparing the score calculated for two addresses, the one that is lower is more likely to be correct (at the particular geography level for which the score was calculated). The address selected with the lowest score for a particular level of geography, is in effect the address most likely to be valid at that level of geography. A detailed description of the development and use of the LRM is provided in Attachment 1 to this document.

For reference purposes, the HUID category definitions are provided in Table 7 below.

Table 7. HUID Category Definitions

Category	Address Type	Definition
A	**2,5,6,A**	Standardized Street Address with TIGER ID, house number and street name
B	**4,7,B**	Property Address with a TIGER ID present
C	**9**	Undefined Address with a TIGER ID present
D	**2,5,6,A**	Standardized Street Address with MAFID, house number and street name
E	**4,7,B**	Property Address with a MAFID present
F	**3 or 8**	P.O. Box or Rural Route Address with MAFID present
G	**9**	Undefined address with a MAFID present
H	**2,5,6,A**	Standardized Street Address with house number and street name – not geocoded
I	**4,7,B**	Property Address, not geocoded
J	**3 or 8**	P.O. Box or Rural Route address – not geocoded
K	**9**	Undefined address
L	**1**	Non-Standardized Address (address standardizer returns all blanks)
M	**n/a**	Bad Address File
Z	**n/a**	Addresses with "Dummy" HUID from the '000' 3-digit ZIP Code File

Note: HUIDs with geocoded data include categories A – G only. All other HUID categories reflect non-geocoded addresses.

1. If only one Linked Person Record exists for a given SSN, output the address record (*HUID*) to the CPR, set the *HUIDSRC* value, and move the requisite geography variables from the Linked Person record to the CPR.
2. If more than one Linked Person Record is present, retain the HUID temporary array until the following processing actions are completed:
   a. Read-in the LRM parameter table (Table A-1 in attachment 1).
   b. Run the LRM equation on each unique HUID within the array to determine the score.
   c. Select the address record with the lowest score for output to the CPR.
   d. If the selected address is from *HUID* category **A – H**, loop through the records in the temporary array (excluding the selected record) to determine the existence of substructure information within *HUID* categories equal to the selected record (if any). If an equal HUID contains substructure information where the selected HUID does not, copy the substructure information into the selected record for output to the CPR.
3. Once the HUID is selected, loop back through the temporary array to determine the source values (*HUIDSRC*). Set the agree/disagree values for each source file based on an exact match of the entire *HUID*.
4. Move the geography variables from the Linked Person File associated with the best HUID to the CPR output file (see the record layout in Section 4).

Create an Initial Comprehensive Address File (ICAF) to house the selected HUID record output to the CPR **and all** other HUID address records for a given SSN. The ICAF (file name is sorted by SSN and retained in 20 segments. The record layout follows:

Table 8. ICAF Record Layout

Label	Description	Type	Length
SSN	Social Security Number	Char	9
UID16	Unique Record Identifier	Char	16
UID17A	Address Residence Type Indicator	Char	1
AID	Address Record Identifier	Char	11
MAFID	Master Address File Identifier Number	Char	9
HUID	Admin Record Housing Unit Identifier	Char	35
SELHUID	Selected HUID (retained on CPR) / Values include: 0 = Not Selected for CPR 1 = Selected HUID for CPR	Char	1
PROXYFLG	Proxy Address Indicator	Char	1
BSACOD	Commercial Flag (from ABI File)	Char	1
BSASRC	Commercial BSA Source Code	Char	3
USPSCD	U.S. Postal Service Record Type Code	Char	1
ATHOME	Work At Home Flag (from ABI File)	Char	1

Note: The number of records written to the ICAF will not equal the total number of records read in for CPR processing as the maximum array for "SSN by group" processing was set to 404 records for a given SSN.

5. Create a Crosswalk File to enable name and SSN matching capability to the CPR. The Crosswalk File will contain the variables displayed in Table 9, and shall be sorted by PIK.

Table 9. Crosswalk File Record Layout

Label	Description	Type	Len
PIK	Protected Identification Key	Char	9
SSN	Social Security Number	Char	9
ARFNM	ADREC First Name	Char	15
ARFNMO	ADREC First Name Overflow Flag **0** = No Character Overflow **1** = Additional Characters in Name Field	Char	1
ARMNM	ADREC Middle Name	Char	15
ARMNM2	ADREC Second Middle Name	Char	15
ARLNM	ADREC Last Name	Char	20
ARLNMO	ADREC Last Name Overflow Flag **0** = No Character Overflow **1** = Additional Characters in Name Field	Char	1

Table 10. Crosswalk File Record Layout (cont.)

Label	Description	Type	Len
ARSUFFIX	ADREC Name Suffix Generation Flag Equivalent (from the name standardizer) Generation Flag Suffix (output value) Blank JR III IV SR	Char	3
STFNM	Standardized First Name	Char	15
STMNM	Standardized Middle Name	Char	15
STMNM2	Standardized Second Middle Name	Char	15
STLNM	Standardized Last Name	Char	20
NMSCORE	Name Score (Calculated from ADREC name) -- presence of a value in the indicated fields will yield a calculated score as follows: Suffix = + 1 Middle Initial = + 2 Middle Full = + 6 First Initial = + 8 First Full = + 24 Last Name = + 32 (maximum name score value = 63)	Char	2
NAMESRC	Name Source (conflict flag) Position: 1 PIC 2 1040 3 1099 4 MEDB 5 CNUM 6 IHS 7 SSS 8 TRACS Value Definition (relative to the selected Name) **0** = No record read for this source **1** = All records agree within source **2** = All non-blank records disagree within source **3** = Conflict within source (2 + 1) **4** = All input records blank within source **5** = Within source records agree and blank (4 + 1) **6** = Within source records disagree and blank (4 + 2) **7** = Within source agree, disagree, and blank (4 + 2 +1)	Char	8
NMDECFL	Selected Name Decision Flag **1** = Highest Name Score **2** = First record read-in among name score ties	Char	1

3.4.3 Miscellaneous Output Variables

A. Protected Identification Key (PIK)

Move the PIK value reflected in the PCF for the current SSN to the PIK field on the CPR.

B. Administrative Record Source (ARSOURCE)

This field provides a tally of all source records that reflect the current SSN/PIK. A value of **9** in the field indicates nine *or more* tallies of the person record.

C. HUID Source Code (HSRC23)

Set a Housing Source Conflict Flag (*HSRC23*) based on an exact match of the first 23 characters (only) of all HUIDs within the array. The agree/disagree values are set in the same fashion as the *HUIDSRC* variable. However, the *HSRC23* values in position 4 (for Numident values) will remain blank for all records. The value output to each field designates the agreement/disagreement among the source records *relative to* the selected demographic variable output to the CPR.

D. Primary Person Record Flag (PRIM)

If no IRS 1040 record is available for the current SSN, retain the *PRIM* default value of zero (**0**). If only one IRS 1040 record exists for the current SSN, move the *PRIM* value from the IRS record to the CPR *PRIM* output variable. If multiple IRS 1040 records exist for the current SSN, tally the bits set for the *PRIM* value to reflect the possible combinations reflected among all the IRS 1040 records (maximum bit set value is 7 [see the record layout in Section 4].

E. Deceased Indicator Type (DCSTYP)

Prior to recalculating a composite value from all the source records, check for a non-blank date-of death in the PCF to determine the value of the "PD" bit in the Deceased Indicator Type (*DCSTYP*) field. For each SSN record check the value in the PCF and complete the processing actions described

1. If the date of death field is non-blank (or the *DODDECFL* assigned value is = **4**), set the *DCSTYP* 'PD' bit to **1**. Otherwise, set the PD bit to **0** (zero).
2. Once the PD bit is set, recalculate a composite value from all source records based on recomputed bit tallies for the entire SSN group (array). Calculate the bit set as 2^{n-1} (maximum value of 15) with each bit defined as follows (see record layout in Section 4):

   0 = Numident Deceased Variable (NV bit)
   1 = PCF Actual Date of Death (PD bit)
   2 = ADREC Deceased Variable (AV bit)
   3 = ADREC Name Standardization (AN bit)

3.4.4 Final Output – Composite Person Record

Once the demographic and geographic (address) selection rules have been invoked and the "best" values determined for retention on the CPR, remove the SSN and all name related fields (shown below) from the output record.

CPR Name related fields (deleted from CPR final)

ARFNM	ARFNMO	ARMNM	ARMNM2	ARLNM	ARLNMO	ARSUFFIX
STFNM	STMNM	STMNM2	STLNM	NMSCORE	NAMESRC	NMDECFL

Within each segment, resort the file by the PIK and output a counts file that includes the following variables: (file name: **`$CNTS08/pbxx_stars08cpr.sas7bdat`**). The CPR record layout is displayed in Section 4.

Counts File Variables

Variable	Label	Type	Length
qcut	Cut number	Character	2
incnt	Number of input records read	Numeric	8
ssncnt	Number of SSNs processed	Numeric	8
lastssn	Last SSN processed	Character	9
outcnt	Number of output records written	Numeric	8
st_time	SAS start time	Numeric	8
en_time	SAS end time	Numeric	8

3.5 Program 38 – Create Linked Unverified Person File

SSNs not verified or found in search during the SSN Search and Verification process are retained in a 21st segment (segment 00), and are not eligible for use in creation of the CPR. Although the unverified person records are not used in creation of the CPR, the records are re-linked to the correct linked address records using the same general processes described for creating the Final Linked Person Files. Thus, the unverified SSN segment undergoes all process actions described in this specification up to creation of the CPR (to include splitting the segment into original source cut files for linkage to the Linked Address Files "LAF"). A Linked Person File for the unverified SSNs is retained in a pre-collapsed form, i.e., no demographic or geographic selection rules are applied to pick a "best record" for an unverified SSN. The DID Modeling and Application Staff may access the pre-collapsed file for evaluation and research purposes.

1. Using program 24, split the records back to the original source file cuts and sort by *UID16* just as the verified person records were. A counts file is also created with the following variables:

Variable	Description	Type	Length
Source	File source identifier (H, I, J, M, T,V, or W)	Char	1
Cut	Cut number with the source file	Char	3
Count	Number of records output to the Unverified Person file per cut	Numeric	8

2. The Final Unverified Person files are then merged by UID16 with the LAF by source in appropriate program (see Process Flow for Program 38) to create the Linked Unverified Person File and a final counts file. The Linked Unverified Person File is retained in only one segment and a final counts file is created with the following variables:

Variable	Description	Type	Length
fid	File source identifier (H, I, J, M, T, V, or W)	Char	1
mcut	Cut number with each source file	Char	3
mrgcnt	Number of records read-in from the Unverified Person File	Numeric	8
addrcnt	Number of records read-in from the Linked Address File	Numeric	8
recout	Number of records output to the Unverified Person File per cut	Numeric	8
perscnt	Number of unmerged records from the Unverified Person File[1]	Numeric	8

Note 1: The number of unmerged records should equal zero (0).

3. Note that during the final merge, each record on the Final Unverified Person File will be matched to at least one record from the LAF. However, not all records on the LAF will match to a record on the Final Unverified Person File.
4. The Linked Unverified Person File is in the same record layout format as the Linked Person File record layout found in Table 2. A visual display of the process may be viewed in the DID-DM StARS Projects process flow area.

4 Output

4.1 Composite Person Record

The Composite Person Record layout is presented in Table 10. The 219-character record represents a single entry, by PIK, for each record holder identified within any of the seven original administrative record source files. The data displayed for each PIK reflects the best demographic and geographic data available for each record based on the selection and unduplication rules defined in the previous sections of this specification.

Note: Conditions often exist where a demographic variable, such as *RACE*, reflects a non-blank value and the source agree/disagree variable (*RACESRC*), reflects all 0 (zero) or four (4) values in the source columns. This condition arises when the PCF must be used to obtain modeled data based on certain selection criteria rules.

Table 10. Composite Person Record Layout

#	Variable	Description	Type	Len
1.	PIK	Protected Identification Key (Replaces Social Security Number)	Char	9
2.	ARSOURCE	Administrative Record Source (Tally Valid Value = 0 – 9 (where 9 = 9 or more occurrences) Position 1 = HUD PIC file Position 2 = IRS 1040 file Position 3 = IRS 1099 file Position 4 = Medicare file Position 5 = Indian Health Services file Position 6 = Selective Service System file Position 7 = HUD TRACS file	Char	7
3.	HUID	Administrative Record Housing Unit Identifier (Category constructed as follows) **A** / FIPST2K / FIPCTY2K / TIGERID / SIDEID / STHN / SUBSTRUCTURE **B** / FIPST2K / FIPCTY2K / TIGERID / SIDEID / STSTNAME **C** / FIPST2K / FIPCTY2K / TIGERID / SIDEID / Sequence # / Blank Fill **D** / FIPST2K / FIPCTY2K / MAFID 9 / Blanks 2 / STHN / SUBSTRUCTURE **E** / FIPST2K / FIPCTY2K / MAFID 9 / Blanks 2 / STSTNAME (blank fill & left justify) **F** / FIPST2K / FIPCTY2K / MAFID 9 / Blanks 2 / STRRDSC / STRRID / STBOXID / Blank Fill **G** / FIPST2K / FIPCTY2K / MAFID 9/ Blanks 2 / Sequence # / Blank Fill **H** / FIPST2K / FIPCTY2K / ZIP5 / Sequence # / STHN / SUBSTRUCTURE **I** / FIPST2K / FIPCTY2K / ZIP5 / Sequence # / STSTNAME (blank fill and left justify) **J** / FIPST2K / FIPCTY2K / ZIP5 / Sequence # / STRRDSC / STRRID / STBOXID / Blank Fill **K** / FIPST2K / FIPCTY2K / ZIP5 / Sequence # / Blank Fill **L** / FIPST2K / FIPCTY2K / ZIP5 / Sequence # / Blank Fill **M** / FIPST2K / FIPCTY2K / ZIP5 / Blank Fill **Z** / FIPST / 999 / 00000 / Sequence # (12) / Blank Fill (12)	Char	35

Table 11. Composite Person Record Layout (cont).

#	Variable	Description	Type	Len
4.	HUIDSRC	Administrative Record Source of HUID Position: 1 PIC, 2 1040, 3 1099, 4 MEDB, 5 CNUM, 6 IHS, 7 SSS, 8 TRACS Value Definition (relative to the selected HUID) **0** = Address not present within source for this person record **1** = All records agree within source **2** = All non-blank records disagree within source **3** = Conflict within source (2 + 1) **4** = All input records blank within source **5** = Within source records agree and blank (4 + 1) **6** = Within source records disagree and blank (4 + 2) **7** = Within source agree, disagree, and blank (4 + 2 +1)	Char	8
5.	HSRC23	Basic Street Address Conflict Flag Position: 1 PIC, 2 1040, 3 1099, 4 MEDB, 5 CNUM, 6 IHS, 7 SSS, 8 TRACS Value Definition (relative to the selected HUID) **0** = Address not present within source for this person record **1** = All records agree within source **2** = All non-blank records disagree within source **3** = Conflict within source (2 + 1) **4** = All input records blank within source **5** = Within source records agree and blank (4 + 1) **6** = Within source records disagree and blank (4 + 2) **7** = Within source agree, disagree, and blank (4 + 2 +1)	Char	8
6.	MAFID	Master Address File Identifier	Char	9
7.	FIPST2K	Census 2000 FIPS Tabulation State Code	Char	2
8.	FIPCTY2K	Census 2000 FIPS Tabulation County Code	Char	3
9.	TRACT2K	Census 2000 Tabulation Tract	Char	6
10.	BLOCK2K	Census 2000 Tabulation Block (includes block suffix)	Char	5
11.	FIPST	Legal (higher) Geography FIPS State Code	Char	2
12.	FPCTY	Legal (higher) Geography FIPS County Code	Char	3
13.	MCD	Current MCD/CCD code	Char	3
14.	PLACE	Current Place Code	Char	4
15.	AIANA	American Indian/Alaskan Native Code	Char	4
16.	CTST	Current FIPS Tabulation State Code	Char	2
17.	CTCTY	Current FIPS Tabulation County Code	Char	3
18.	TRACT	Current Tabulation Tract — Current TIGER / Current MAF Values	Char	6
19.	BLOCK	Current Tabulation Block	Char	4
20.	BLKSFX	Current Tabulation Block Suffix	Char	1

Table 11. Composite Person Record Layout (cont).

#	Variable	Description	Type	Len
21.	ZIP53	ZIP Code (first 3 digits of 5-digit ZIP)	Char	3
22.	ZIP52	ZIP Code (last2 digits of the 5-digit ZIP)	Char	2
23.	ZIP4	ZIP Code + 4 value	Char	4
24.	BSACOD	Commercial Flag (from ABI File)Value: Definition **0** = Residential – known single unit **1** = Residential – possible multi-unit **2** = Apartment Buildings **3** = Hotels/Motels **4** = Mobile home parks/marinas/RV parks & campsites **5** = Group Quarters (excluding Hotels and Motels) **6** = Commercial – Business Address single unit **7** = Commercial – Business Address multi-unit **8** = Mixed Use – Doctors/Lawyers/Real Estate Offices, etc. **9** = Mixed Use – other than type 8 **A** = Unmatched to Commercial Address	Char	1
25.	BSASRC	Commercial BSA Source Code (bit set to indicate the various types of commercial addresses present at a given basic street address) Bit Set Flag — Definition **0** — Apartment Buildings **1** — Hotels/Motels **2** — Mobile home parks/marinas/RV parks & campsites **3** — Group Quarters (excluding Hotels and Motels **4** — Commercial – Business Address single unit **5** — Commercial – Business Address multi-unit **6** — Mixed Use – Doctors/Lawyers/Real Estate Offices, etc. **7** — Mixed Use – other than type 8 **8** — Unmatched to Commercial Address	Char	3
26.	ATHOME	Work At Home Flag **Blank** = Default (not flagged as a work-at-home address) **1** = Identified as a work at home address	Char	1
27.	PROXYFLG	Proxy Address Indicator	Char	1
28.	PRIM	Record of Primary Person Flag Value — Bit Set — Meaning **0** — none — NA (default setting) **1** — 0 — Primary Person(s) **2** — 1 — Secondary Person(s) **3** — 0,1 — Primary/Secondary Persons Combination **4** — 2 — Dependent Person(s) **5** — 0,2 — Primary/Dependent Persons Combination **6** — 1,2 — Secondary/Dependent Persons Combination **7** — 0,1,2 — Primary/Secondary Dependent Persons Combination	Char	1
29.	SEX	Gender (Blank = No Data Present / not applicable) **1** = Male **2** = Female	Char	1

Table 11. Composite Person Record Layout (cont).

#	Variable	Description	Type	Len
30.	SEXSRC	Gender Source Position: 1 PIC, 2 1040, 3 1099, 4 MEDB, 5 CNUM, 6 IHS, 7 SSS, 8 TRACS Value Definition (relative to the selected Sex) **0** = No record read for this source **1** = All records agree within source **2** = All non-blank records disagree within source **3** = Conflict within source (2 + 1) **4** = All input records blank within source **5** = Within source records agree and blank (4 + 1) **6** = Within source records disagree and blank (4 + 2) **7** = Within source agree, disagree, and blank (4 + 2 +1)	Char	8
31.	SEXDECFL	Selected Gender Decision Flag Selection Rule Invoked: **1** = Male if record appears on Selective Service **2** = Most frequent observation **3** = PCF value	Char	1
32.	BESTRACE	Best Last Race (from Census Numident) Blank = Missing **0** = Unknown **1** = White **2** = Black **3** = Other **4** = Asian and Pacific Islander **5** = Hispanic **6** = North American Indian or Eskimo **7** = Reserved for future use **8** = Reserved for future use	Char	1
33.	ARRACE	Administrative Record Race (Blank = No Data Present / not applicable) **1** = White **2** = Black **3** = American Indian, Eskimo, or Aleut **4** = Asian or Pacific Islander	Char	1
34.	RACESRC	Race Source Conflict flag Position: 1 PIC, 2 1040, 3 1099, 4 MEDB, 5 CNUM, 6 IHS, 7 SSS, 8 TRACS Value Definition (relative to the selected Race) **0** = No record read for this source **1** = All records agree within source **2** = All non-blank records disagree within source **3** = Conflict within source (2 + 1) **4** = All input records blank within source **5** = Within source records agree and blank (4 + 1) **6** = Within source records disagree and blank (4 + 2) **7** = Within source agree, disagree, and blank (4 + 2 +1)	Char	8
35.	RACDECFL	Selected Race Decision Flag (selection rule invoked) **1** = AI/AN flag from Indian Health Service File **2** = Most Frequent Observation **3** = No administrative record data available (blank) or ties among most frequent observation **4** = PCF Census Race **5** = PCF Modeled Race	Char	1

Table 11. Composite Person Record Layout (cont).

#	Variable	Description	Type	Len
36.	LKLY_RACE	Most likely race from Census Race Response or Race Model **W** = White **B** = Black **I** = Indian **A** = Asian **P** = Pacific Islander **M** = Multiple Race	Char	1
37.	CERADEFL	Census Race Decision Flag **1** = Census Race (from HDEF) **2** = Modeled Race	Char	1
38.	ARHISP	Administrative Record Hispanic Origin // Blank = No Data Present / not applicable **1** = Hispanic **2** = Not Hispanic	Char	1
39.	HISPSRC	Hispanic Source Position: 1 PIC; 2 1040; 3 1099; 4 MEDB; 5 CNUM; 6 IHS; 7 SSS; 8 TRACS Value Definition (relative to the selected Hispanic Origin) **0** = No record read for this source **1** = All records agree within source **2** = All non-blank records disagree within source **3** = Conflict within source (2 + 1) **4** = All input records blank within source **5** = Within source records agree and blank (4 + 1) **6** = Within source records disagree and blank (4 + 2) **7** = Within source agree, disagree, and blank (4 + 2 +1)	Char	8
40.	HISDECFL	Selected Hispanic Decision flag (Selection Rule Invoked) – blank = no data present / not applicable **1** = Most Frequent Observation **2** = No admin record data available (blank) or ties among most frequent observation **3** = PCF Census Hispanic Origin **4** = PCF Modeled Hispanic Origin	Char	1
41.	LKLY_HISP	Most likely Hispanic Origin from Census Race Response or Race Model **Y** = Hispanic **N** = Non-Hispanic	Char	1
42.	CHSPDEFL	Census Hispanic Origin Decision Flag **1** = Census (HDEF) Response **2** = Modeled Value	Char	1
43.	CYDOB	Birth Date (Century and Year) Blank = no data present or N/A CCYY format – valid range between 1886 – 2008	Char	4
44.	MMDOB	Birth Date (Month) Blank = no data present or N/A MM format – Valid range between 01 – 12	Char	2
45.	DDDOB	Birth Date (Day) Blank = no data present / not applicable DD format – valid range between 01 -31	Char	2
46.	DOBSCOR	Date of Birth Score (* = non blank // - = blank) YYYY MM DD Score: * * * 14; * * - 12; * - * 10; * - - 08; - * * 06; - * - 04; - - * 02; - - - 00	Char	2

Table 11. Composite Person Record Layout (cont).

#	Variable	Description	Type	Len
47.	DOBSRC	Date of Birth Source (conflict flag) Position: 1 PIC, 2 1040, 3 1099, 4 MEDB, 5 CNUM, 6 IHS, 7 SSS, 8 TRACS Value Definition (relative to the selected DOB) **0** = No record read for this source **1** = All records agree within source **2** = All non-blank records disagree within source **3** = Conflict within source (2 + 1) **4** = All input records blank within source **5** = Within source records agree and blank (4 + 1) **6** = Within source records disagree and blank (4 + 2) **7** = Within source agree, disagree, and blank (4 + 2 +1)	Char	8
48.	DOBDECFL	Date of Birth Decision Flag // Selection Rule Invoked: **1** = Selected from Medicare **2** = From Selective Service Record **3** = Census Numident **4** = HUD TRACS **5** = HUD PIC **6** = Indian Health Service **7** = No date of birth available (blank on CPR)	Char	1
49.	CYDOD	Date of Death (Century and Year) Blank = No Data Present / not applicable CCYY format – valid range between 1886 – 2008 (inclusive)	Char	4
50.	MMDOD	Date of Death (Month) Blank = No Data Present / not applicable MM format – valid range between 01 –12	Char	2
51.	DDDOD	Date of Death (Day) Blank = No Data Present / not applicable DD format – valid range between 01 -31	Char	2
52.	DODSRC	Date of Death Source Position: 1 PIC, 2 1040, 3 1099, 4 MEDB, 5 CNUM, 6 IHS, 7 SSS, 8 TRACS Value Definition (relative to the selected DOD) **0** = No record read for this source **1** = All records agree within source **2** = All non-blank records disagree within source **3** = Conflict within source (2 + 1) **4** = All input records blank within source **5** = Within source records agree and blank (4 + 1) **6** = Within source records disagree and blank (4 + 2) **7** = Within source agree, disagree, and blank (4 + 2 +1)	Char	8

Table 11. Composite Person Record Layout (cont).

#	Variable	Description	Type	Len
53.	DODDECFL	Selected Date of Death Decision Flag (Selection Rule Invoked) **1** = Medicare DOD Record **2** = Most Recent Medicare Record **3** = Numident DOD Record **4** = PCF reflects date of death information	Char	1
54.	DCSTYP	Deceased Indicator Type **-** = blank (no data present) ***** = indication person is deceased from the following sources: AN = ADREC Name Standardization AV = ADREC Deceased Variable PD = PCF Actual Date of Death NV = Numident Deceased Variable AN AV PD NV Bits Set Value * * * * 0,1,2,3 15 * * * - 1,2,3 14 * * - * 0,2,3 13 * * - - 2,3 12 * - * * 0,1,3 11 * - * - 1,3 10 * - - * 0,3 09 * - - - 3 08 - * * * 0,1,2 07 - * * - 1,2 06 - * - * 0,2 05 - * - - 2 04 - - * * 0,1 03 - - * - 1 02 - - - * 0 01 - - - - 00 Calculate the bit set as 2^{n-1} (maximum value of 15) with each bit defined as follows: Bit Set Deceased Indicator Source Definition **0** Numident Deceased Variable (NV bit) **1** PCF Actual Date of Death (PD bit) **2** ADREC Deceased Variable (AV bit) **3** ADREC Name Standardization (AN bit)	Char	2
55.	VERTYPE	SSN Verification Type Code Calculate the bit set as 2^{n-1} (maximum value of 63) with each bit defined as follows: Bit Set Meaning **0** = Verified **1** = Verified – IRS 1040 Special Rules **2** = Verified – IRS 1099 Special Rules **3** = Verified – IRS 1099 child/parent name switch **4** = Found in GEOKEY Search **5** = Found in Name Search	Char	2

5 QA Requirements

The following are general Quality Control (QC) checks required for both the address (EAF) and person (EPF) output files. The checks are extracted from the DID Program Review and Quality Control modular procedures document. QC will be completed in accordance with the instructions found within the Quality Control Procedures Document for each of the modules specified below by populating the date column of the QC Test Module form to signify completion of the given QC line item. For each line item, the QC analyst will enter a "**YES**" or "**NO**" in the "Result" column of the test module form to indicate QC compliance. Most "**NO**" responses require QC analyst comments in the box following the "Results/Date" column on the appropriate QC Test module document.

5.1 Module 1 – Program Log File Review

The log file review requires a check of the Program Log to check for error, warnings and unexpected note messages for indications of an "unclean" run. Methods of reviewing the SAS log file may vary, but one way is to:

1. Read the log into the Pico editor (type "pico *logname.log*" on the command line).
2. Hold the "Ctrl" button down at the same time as you type "W" – this is a pico search command.
3. Then type the word you are looking for: for example, type *error.*
4. Pico will either return "error not found" or will go to the place in the log where that word is found.
5. The Pico editor is ***not*** case sensitive.

5.2 Module 2 – Record Count Review

Complete a program input/output record count review using the example of a SAS macro that can read in 20 file cuts, append the number of observations of each cut into a single SAS dataset, and print the results. To change the number of cuts read in, change the "do" statement. The code below is designed to read in cuts 01 to 20, with a zero formatted into the cut numbers. Change the "z" in the put statement to "3" (so it reads "z3." if you wish to pad two zeros in front of the cut number.

```
Libname yourlib "/yourdir/yourdir";
%macro younameit;
      %do i = 1 %to 20;
      data _null_;
            call symput ('cut',put(&i,z2.));
            ***'i'=iterations'z2'=cut # format ***;
      run;
      proc contents data=yourlib.yourdataset&cut out=a
                  (keep = nobs) noprint;
      run;
      data b;
            set a;
            if _n_ = 1;  ***starts at the first observation ***;
```

```
            cut = &cut.; ***names the cut # "cut"          ***;
      run;
      *** The append creates one dataset     ***;
      *** with the # obs from each iteration ***;
      proc append base = yourlib.appended_dataset data = b;
      run;
      *** you must delete the old appended file ***;
      *** if you re-run the code                ***;
%end;
%mend younameit;
%younameit;
Proc print data = yourlib.appended_dataset;
      Sum nobs;
      ** this will sum up the number of observations **;
      Title1 "THIS IS MY COUNTS FILE";
Run;
```

5.3 Module 3 – Program Code Review

The only Program Code Review required for the Composite Person Record program is to run a UNIX DIFF of the 2008 version of the file against the 2005 version, since program code has not been altered since last year's production run.

5.4 Module 4 – Output File Review

1. Draw a simple random sample of 100,000 records from the output file, and order this sample randomly. Perform an output file review on the random sample using the test modules that follow. A simple random sample is one in which all records have an equal probability of being selected regardless of cut or record characteristics. The following code shows how this can be accomplished.

```
%macro younameit;
   %do i = 1 %to 1;
      data _null_;
      call symput ('cut',put(&i,z2.));
      ***'i'=iterations'z2'=cut # format***;
      run;
      data outlist2;
            set a.pbxx_all_males&cut;
            where ranuni(0) < selection_probability;
             ** see note below;
      run;
      proc append base=alloutlist2 data=outlist2;
      run;
   %end;
%mend;
%younameit;

data alloutlist2;
   set alloutlist2;
   rn = ranuni(0);
run;
proc sort data=alloutlist2;
```

```
   by run;
run;
proc print data=alloutlist2 (obs=200);
   title Random Listing;
run;
```

Note: *selection_probability* is set just large enough to insure that at least 100,000 records will be selected from all of the cuts combined. Thus, *selection_probability = 100000 / total output records.*

If the total number of records in the output file is less than 100,000, set the *selection_probability* = 1.

5.4.1 Module 4A – Free Text Fields

1. Free Text Fields include output values for names, street addresses, cities, etc. These fields should be reviewed for reasonable values from the random sample. For example do first names look like first names, do the street addresses look to be in the proper format? Is the last name field filled for all or most of the records? The determination of a reasonable output value is subject to the judgment of the QC analyst.
2. Review the random sample and enter a **YES**, **NO**, or not applicable (**N/A**) response in the "Result" column of the appropriate Test Module document. A **NO** response requires explanation in the comments box.

5.4.2 Module 4B – Categorical Fields

1. Categorical fields include such variables as edit flags. These fields should be reviewed using PROC FREQ. For each field to be reviewed the PROC FREQ output should be scanned. Does the distribution make sense based on
   - the analysts understanding of the input files and processes,
   - guidance provided by data provider and customer, if any, and
   - frequency distributions produced from previous cycles.
2. Review the PROC FREQ of the categorical fields from the random sample and enter a **YES**, **NO**, or not applicable (**N/A**) response in the result column. Any categorical field output values that caused a **NO** response must be listed in the comments box and an explanation of the deviation provided.

5.4.3 Module 4C – Geocode Fields

1. The Geocode Fields include MAFIDs, TRACT Numbers, TIGERIDs, HUIDs, AIDs, Block codes, etc. These fields should also be reviewed using PROC FREQ. For each field to be reviewed the PROC FREQ output should be scanned, and a frequency produced by major classification categories.
2. Review the PROC FREQ of the geocode fields from the random sample and enter a **YES**, **NO**, or not applicable (**N/A**) response in the result column. Any geocode field output values that caused a **NO** response must be listed in the comments box and an

explanation of the deviation supplied. The QC analyst checks for reasonable distribution, acceptable levels of illegal values, and determines the validity of illegal values based on the documentation.

5.4.4 Module 4E – Date Fields

1. Months and days should be reviewed using PROC FREQ.
   a. Check for illegal values.
   b. Check for reasonable distributions.

If these fields are part of a Date-of-Birth, Date-of-Death, or other unscheduled event, the distribution across the various possible values should be flat.

If the distribution is uneven, the analyst should decide whether this distribution is reasonable given the type of date value that is represented.

2. Year fields should be reviewed with PROC CHART.
   a. If the year value is in two separate fields (i.e. century and year of century), then these values should be concatenated before using PROC CHART.
   b. For each field, check for reasonable values from the PROC CHART output based upon:
      1) understanding of the input files and processes,
      2) any guidance provided by the customer or data provider and,
      3) frequency distributions produced from previous cycles.

Review the PROC FREQ and PROC CHART of the date fields from the random sample and enter a **YES**, **NO**, or not applicable (**N/A**) response in the result column. Any date field output values that caused a **NO** response must be listed in the comments box (on the appropriate Test Module Form) and an explanation of the deviation provided.

5.4.5 Module 4F – SSN & ID Number Fields

The following code can detect, count and output illegal SSNs or other key identification variables to a separate file for review. What constitutes an "illegal code" will vary on the project, so that should be determined by the project team members.

1. Using the random sample output file taken previously, run the following code:

```
/* Notes -- randsamp = your random sample SAS dataset
            idnum = the SSN or other ID field
*/
proc sort data=randsamp (keep=idnum) out=temp;
      by idnum;
run;
data temp;
      set temp end=last;
      by idnum;
      if idnum in (illegal codes) then illegal_count + 1;
            ***illegal codes to be determined****;
      if first.idnum then count = 1;
```

```
            else count + 1;
      if last.idnum then output;
      if last then put illegal_count=;
run;
proc sort data=temp;
      by descending count;
run;
proc print data=temp (obs=200);
      title Most Frequently Appearing ID Numbers;
      var idnum count;
run;
```

2. The QC analyst will review the listing to determine reasonable values based upon:
    - The analysts' understanding of the file
    - Past experience with this type of file
    - Guidance (if any) provided by the customer and data provider.

6 Change History

Date Change Made	Author	Authorizer	Change
12/20/2008	Mercedes Butler	DPB	**Removed specific File and Program name references. All file and program names documented through the Process Flow associated with the specification name.**

Attachment 1. Logistic Regression Model — StARS Address Selection

The information presented in the attachment is the result of an analysis completed on the StARS 2000 address quality and selection by Dean Resnick, a member of the DID Data Analysis staff. Results of the evaluation led to development and employment of a logistic regression model to select the best address for retention on the StARS 2008 Composite Person Records.

Administrative address (StARS CPR address) correctness determined by comparison to Census 2000.

To evaluate CPR address correctness, DID staff compared Linked Person File (StARS 2000) addresses to addresses in the Census Hundred Percent Detail Edited File (HDEF). The comparison was accomplished by determining which HDEF record belonged to each person in the CPR (a StARS file derived by summarizing the Linked Person File). The DID Race Enhanced Numident Project required that HDEF records be linked with Numident records. From this project, DID possessed an index file showing for each (but not every) Social Security Number (SSN), the likely HDEF record for the same person. As the Linked Person File is indexed by SSN, it is not difficult to associate the various Linked Person File addresses and the HDEF record for a person. There were, of course, persons in the HDEF for whom the SSN could not be determined, and there are also persons in the Linked Person File, for whom an HDEF record could not be found. The data for such persons was not used in this evaluation. Persons for whom matched records were obtained, a simple random sample was selected at a rate of one record per hundred.

The next part of the analysis compared addresses in the Linked Person file to the HDEF address for each person in the sample. Comparisons were made at several levels of geography: state, county, tract, block, and MAFID. These geography elements were in the format used for Census 2000. Several statistics to describe how well the HDEF address matched those in the Linked Person File were tabulated. The statistics were generated for each of the selected geography elements:

		State	County	Tract	Block	MAFID
1.	Persons in Analysis	2,374,450	2,374,450	2,380,153	2,380,153	2,374,450
2.	# w/ correct choice avail.	2,264,490	2,205,179	1,973,756	1,925,725	1,784,432
	%. of Above (2 ÷ 1)	95.4%	92.9%	82.9%	80.9%	75.2%
3.	Persons w/ 1 choice	2,281,811	2,181,232	2,064,385	2,027,986	1,992,574
	%. of Total (3 ÷ 1)	96.1%	91.9%	86.7%	85.2%	83.9%
4.	# w/ correct choice	2,173,578	2,019,910	1,687,525	1,615,869	1,464,391
	%. of Above (4 ÷ 3)	95.3%	92.6%	81.7%	79.7%	73.5%
5.	Person w/ > 1 choice	92,639	193,218	315,768	352,167	381,876
	%. of Total (5 ÷ 1)	3.9%	8.1%	13.3%	14.8%	16.1%
6.	# w/ correct choice avail.	90,912	185,269	286,231	309,856	320,041
	%. of Total (6 ÷ 5)	98.1%	95.9%	90.6%	88.0%	83.8%

From the above data, it is apparent that moving from the highest level of geography (state) to the lowest level of geography (MAFID), the percentage of persons for whom more than one location is available from administrative records increases. Also the percentage of persons for whom the correct location is available for selection from within administrative records decreases, both for those with only one location available for selection and those with more than one location available for selection. Across all levels of geography, the greater the percentage of persons with more that one choice of location from administrative records, the greater the importance of good address selection.

Analysis continued by attempting to answer the following questions:

- *How good are DID current procedures for StARS address selection?*
- *Can DID develop an alternate procedure for address selection that improves the likelihood that the best location from among those available from administrative records for each person is selected?*

The results gathered from the second question enables a presentation of data that answers the first question later in the paper. To answer the second question, an alternate procedure was built based on the concept that for each person, each unique address seen for this person has a likelihood of being correct, and this likelihood of being correct differs for the various levels of geography that locate this address. For example, for a person with two unique addresses seen within the administrative records, each address would have a likelihood of being correct at the state level, a likelihood of being correct at the county level, a likelihood of being correct at the tract level, a likelihood of being correct at the block level, and a likelihood of being correct at the MAFID level. The problem was conceived in this way for maximum flexibility and to allow for cases where one of the addresses does not have complete geography. In such cases, even though an address may be location specific only for state and county, this state and county may be the correct one, whereas a more specific address may be incorrect for state and county or for just state.

To model the probabilities that unique addresses were correct, logistic regression modeling was used. For this technique, probability is modeled as a function of several indicator variables. In order to implement this technique, possible indicators of address validity were identified and tested within logistic regression. The following indicators were initially tested:

Variable	Description	Variable	Description
geosoflg	Geography source flag	source file	Admin record file source indicator
cenaddr	Good Census address	msf	multiple source file records for the same address
addrtype	HUID address category type	un_av	whether unit data is available on street address
geotype	Type of geocoding flag	race	race
bsacod	Commercial flag from ABI file	sex	gender
athome	Work at home flag	Hisp	Hispanic origin
proxyflg	Proxy address flag indicator	age	age
uspscd	USPS state code	cycle date	Only PIC, IRS1040, MEDB, SSS, & IHS records

Only the indicators in the following table were actually found useful for determining address validity, and were redefined in order the make the variables usable for logistic regression modeling as indicated:

Variable	Redefined Description
source file	Separated into indicator for each source file as a binary value where: fs_h = PIC fs_i = IRS 1040 fs_j = IRS 1099 fs_m = MEDB fs_t = IHS fs_v = SSS fs_w = HUD TRACS Example: if a person's unique address had a source of IRS-1040 and Medicare only, then fs_i = fs_m =1 and fs_h = fs_j = fs_t = fs_v = fs_w = 0.
bsacod	Separated into indicators for each code. Only the following values were found significant: cf_3 = commercial flag set to 5 (group quarters) and cf_5 = commercial flag set to 7 (commercial)
proxyflg	Npf_1 where the proxy flag was not set to 1 (indicating ***not*** a proxy address) on the record
cycle date	Cycle dates converted to displacement days (# of days prior to 4/2/2000) expressed as a negative number. One indicator variable created for each of the useful files as follows: did = PIC cycle date dd_i = IRS 1040 cycle date dd_y = Selective Service cycle date
athome	ah indicate at home flag set as follows: 0 = default (no work at home), and 1 = work at home
addrtype	HUID address category type separated into indicators for each address type as follows: hat_a = HUID category A hat_b = HUID category B hat_c = HUID category C (all categories A through M included)
msf	Multiple source file indicator (set to a binary value)
un_av	Unit data (substructure data) available for the record – also set to a binary value

All of these redefined indicators except cycle date (dd_h, dd_i, and dd_v) were set to binary values (i.e., 0 or 1 depending on whether the specific condition were true). From these recoded variables, various interaction terms were built and tested, with the following interaction terms surviving testing and incorporated into the final model:

```
npfcf3 = npf_1 x cf_3          npfcf5 = npf_1 x cf_5
fsih = fs_i x fs_h             fsij = fs_i x fs_j
fsim = fs_i x fs_m             fsit = fs_i x fs_t
fsiv = fs_i x fs_v             fsiw = fs_i x fs_w
fsjh = fs_j x fs_h             fsjm = fs_j x fs_m
fsjt = fs_j x fs_t             fsjv = fs_j x fs_v
fsjw = fs_j x fs_w             fsijm = fs_i x fs_j x fs_m
fsicf3 = fs_i x cf_3           fsjcf3 = fs_j x cf_3
fsmcf3 = fs_m x cf_3           fsicf5 = fs_i x cf_5
fsjcf5 = fs_j x cf_5           fsmcf5 = fs_m x cf_5
fsinpf1 = fs_i x npf_1         fsjnpf1 = fs_j x npf_1
                    fsmnpf1 = fs_m x npf_1
```

The interaction terms were developed by trying all combinations of the simple variables that potentially seemed able to have explanatory value. Of the interaction terms tested, only those found to be statistically significant (at the 5% confidence level) were retained.

Running the regression model using the indicator variables (both simple and interaction) yielded a set of parameters for each level of geography, one for each indicator variable. Some indicator variables were assigned a parameter value of zero because the estimated value for the parameter was insignificantly different from zero. The following table presents the parameter estimates that were made for each level of geography:

Table A-1. LRM Parameter Table

	Geography Level				
Parameter	***State***	***County***	***Tract***	***Block***	***MAFID***
Intercept	3.19400	0.89720	3.08830	3.36050	3.69220
fs_h	-2.77740	-2.77570	-3.04090	-3.31200	-2.92660
fs_i	-0.34980	-0.25680	-0.93280	-1.42540	-1.82150
fs_j	-0.66260	-0.27520			0.53250
fs_m	-0.69010	-0.54090	-1.36420	-1.80400	-1.52480
fs_t	-0.30860	-0.18250	0.79850	1.02490	1.15640
fs_v	-0.40900	-0.48990	-0.39420	-0.31380	
fs_w	-2.81100	-2.81890	-2.86880	-2.14170	3.64640
dd_h	-0.00501	-0.00407	-0.00175	-0.00200	-0.00210
dd_i	-0.00435	0.00029	-0.00608	-0.00476	-0.00521
dd_v	-0.00047	-0.00057	-0.00080	-0.00088	-0.00084
cf_3		-1.06860	0.41610		0.75180
cf_5	0.14420	1.37150	2.29070	2.83000	3.53760
npf_1		0.18760	-0.54440	-0.80430	-1.11160
msf	-0.28940	-0.25090	-0.39340	-0.54820	-0.81030
ah	0.31070	0.40100	0.65610	0.60880	0.70660
un_av	-0.42820	-0.46260	-0.37290	-0.14050	-0.88000
npfcf3	0.25430	1.36300			
npfcf5		-0.81350	-1.09130	-1.76670	-1.78210
fsih	0.75510	0.93120	1.02900	1.26700	1.39300
fsij		0.03020		0.26610	
fsim	0.66110	0.60640	1.10340	1.44130	1.62430
fsit	-0.40320	-0.32000	-0.74290	-0.35400	-0.49870
fsiv	1.09860	1.12390	0.84510	0.70420	0.59410
fsiw		0.85860	1.98880	1.60860	
fsjh	0.38080	0.49280	1.07410	1.12560	0.70900
fsjm	-0.42280	-0.48440	-0.37700		-0.36060
fsjt	-0.39810	-0.53180	-0.67690	-0.61600	-1.05510
fsjv	0.73440	0.68390	0.48720	0.31200	-0.07680
fsjw		0.74120	0.33920	0.33650	
fsijm		-0.02680	-0.21920	-0.54190	-0.44460
fsicf3		0.13580		0.31160	
fsjcf3		-0.00465		0.31000	
fsmcf3		-0.17950			
fsicf5	0.85800	0.54250	0.49740	0.37340	-0.15480

Table A-1. LRM Parameter Table (cont).

fsjcf5		-0.02720			-0.20150
fsmcf5		-0.63820	-0.73750	-0.83550	-1.01030
fsinpf1	-1.79850	-1.85950	-1.77800	-1.45420	-0.88610
fsjnpf1	-0.23440	-0.55230	-0.80380	-0.90740	-0.85990
fsmnpf1		-0.10960	0.35510	0.50910	0.43400
hat_a	-1.73520	0.45840	-0.33460		
hat_b	-2.56190	0.48330			
hat_c		-6.45700			
hat_d	-1.77100	0.43910	-0.34840		0.73710
hat_e		0.34390	-1.55720		
hat_f	-1.21320	0.77030			0.23470
hat_g	-1.95670	-0.02220			2.47250
hat_h	-2.17540	0.49590			
hat_i	-1.98050	0.37550			
hat_j	-1.10820	1.45100			
hat_k	-2.23070	0.06090			
hat_l	-2.61080	-0.16250			
hat_m	-2.20100	0.00000			

In a logistic regression model, the estimated probability is determined by the equation:

(1) $$probability\ of\ address\ invalidity = \frac{1}{1+e^{-l.e.}}$$

where

(2) $$l.e.\,(linear\ estimator) = \sum_{i=1}^{n} parameter_i \cdot estimator_i$$

and n = the number of estimators

Note that in the above equation, as the linear estimator becomes larger, the estimated probability also becomes larger. This can be restated by saying that when comparing two linear estimators, the larger always generates a larger probability that a smaller. This property of this relationship is often described as one of a monotonically increasing function.

However, somewhat confusingly, because the software estimated the probability of geographical components being incorrect rather than correct, the more negative a parameter is in the above table, the more the associated predictor's existence (or size) increases the probability of an address being valid. Thus, for example, looking at the probability that an address is accurate at the state level, all things being equal an address with a sole source of either PIC (parameter value for fs_h = -2.77740) or TRACS (parameter value for fs_w = -2.81100) is significantly more likely to be correct that an address with a sole source of IRS-1040 (parameter value for fs_i = -0.34980). Interestingly, at the MAFID level, an address with a sole source of IRS-1040 (parameter value for fs_i = -1.82150) is much more likely to be valid than with a sole source of TRACS (parameter value = 3.64640). *In fact, TRACS addresses are indicated to be accurate at all geography levels other than MAFID; perhaps this effect results from "bad" substructure identification for addresses (which are otherwise*

accurate) in the TRACS file. Where cells are left blank on the above table, the parameter value was not found to be statistically significant and the associated predictor was not included in the model. Non-inclusion in the model has the equivalent effect of giving the associated parameter a value of zero.

A score that measures address correctness likelihood is developed with logistic regression. Because of the monotonically increasing property of the logistic regression function, the linear estimator can be considered a score for an address. Where the score is low at a particular level of geography, the address is more likely to be correct at that level. So by comparing the score calculated for two addresses, the one that is lower is more likely to be correct (for the particular geography level the score was calculated). The linear estimator equation shows that the score is equal to the sum of the products of estimator values and their associated parameters. Table A-2 provides a score computation example.

Table A-2. LRM Score Computation Table

Indicator	Value	Comments	Parameter Value	Product
Intercept	1	The intercept value is always included in the total.	3.19400	3.19400
fs_h	0	This address did not have PIC as a source.	-2.77740	0.00000
fs_i	1	This address did have IRS-1040 as a source.	-0.34980	-0.34980
fs_j	1	This address did have IRS-IRMF (1099) as a source.	-0.66260	-0.66260
fs_m	1	This address did have MEDB (Medicare) as a source.	-0.69010	-0.69010
fs_t	0	This address did not have IHS as a source.	-0.30860	0.00000
fs_v	0	This address did not have SSS as a source.	-0.40900	0.00000
fs_w	0	This address did not have TRACS as a source.	-2.81100	0.00000
dd_h	0	Set to 0 because PIC is not a source.	-0.00501	0.00000
dd_i	-14	The IRS-1040 cycle date was 14 days prior to 4/1/2000. That is, 3/18/2000.	-0.00435	0.06090
dd_v	0	Set to 0 because SSS is not a source	-0.00047	0.00000
cf_3	0	Set to 0 because address did not have a commercial flag of 3 (that is, ABI does not show this address as a group quarters).		0.00000
cf_5	0	Set to 0 because address did not have a commercial flag of 5 (that is, ABI does not show it as a commercial address).	0.14420	0.00000
npf_1	1	Set to 1 because PROXYFLG does not equal 1.		0.00000
msf	1	Set to 1 // address present on multiple source files.	-0.28940	-0.28940
ah	0	Set to 0 because athome = 0.	0.31070	0.00000
un_av	1	Set to 1 // unit data (substructure ID) is available.	-0.42820	-0.42820
npfcf3	0	Set to 0 as product of npf_1 = 1 and cf_3 = 0.	0.25430	0.00000
npfcf5	0	Set to 0 as product of npf_1 = 1 and cf_5 = 0.		0.00000
fsih	0	Set to 0 as product of fs_i = 1 and fs_h = 0.	0.75510	0.00000
fsij	1	Set to 1 as product of fs_i = 1 and fs_j = 1.		0.00000
fsim	1	Set to 1 as product of fs_i = 1 and fs_m = 1.	0.66110	0.66110
fsit	0	Set to 0 as product of fs_i = 1 and fs_t = 0.	-0.40320	0.00000
fsiv	0	Set to 0 as product of fs_i = 1 and fs_v = 0.	1.09860	0.00000

fsiw	0	Set to 0 as product of fs_i = 1 and fs_w = 0.		0.00000
fsjh	0	Set to 0 as product of fs_j = 1 and fs_h = 0.	0.38080	0.00000
fsjm	1	Set to 1 as product of fs_j = 1 and fs_m = 1.	-0.42280	-0.42280
fsjt	0	Set to 0 as product of fs_j = 1 and fs_t = 0.	-0.39810	0.00000
fsjv	0	Set to 0 as product of fs_j = 1 and fs_v = 0.	0.73440	0.00000
fsjw	0	Set to 0 as product of fs_j = 1 and fs_w = 0.		0.00000
fsijm	1	Set to 1 as product of fs_i = 1, fs_j = 1, and fs_m = 1.		0.00000
fsicf3	0	Set to 0 as product of fs_i = 1 and cf_3 = 0.		0.00000
fsjcf3	0	Set to 0 as product of fs_j = 1 and cf_3 = 0.		0.00000
fsmcf3	0	Set to 0 as product of fs_m = 1 and cf_3 = 0.		0.00000
fsicf5	0	Set to 0 as product of fs_i = 1 and cf_5 = 0.	0.85800	0.00000
fsjcf5	0	Set to 0 as product of fs_j = 1 and cf_5 = 0.		0.00000
fsmcf5	0	Set to 0 as product of fs_m = 1 and cf_5 = 0.		0.00000
fsinpf1	1	Set to 0 as product of fs_i = 1 and npf_1 = 1.	-1.79850	-1.79850
fsjnpf1	1	Set to 0 as product of fs_j = 1 and npf_1 = 1.	-0.23440	-0.23440
fsmnpf1	1	Set to 0 as product of fs_m = 1 and npf_1 = 1.		0.00000
hat_a	0	Set to 0 as HUID Address Type is not 'A'.	-1.73520	0.00000
hat_b	0	Set to 0 as HUID Address Type is not 'B'.	-2.56190	0.00000
hat_c	0	Set to 0 as HUID Address Type is not 'C'.		0.00000
hat_d	0	Set to 0 as HUID Address Type is not 'D'.	-1.77100	0.00000
hat_e	0	Set to 0 as HUID Address Type is not 'E'.		0.00000
hat_f	0	Set to 0 as HUID Address Type is not 'F'.	-1.21320	0.00000
hat_g	0	Set to 0 as HUID Address Type is not 'G'.	-1.95670	0.00000
hat_h	0	Set to 0 as HUID Address Type is not 'H'.	-2.17540	0.00000
hat_i	0	Set to 0 as HUID Address Type is not 'I'.	-1.98050	0.00000
hat_j	1	Set to 1 as HUID Address Type is 'J'.	-1.10820	-1.10820
hat_k	0	Set to 0 as HUID Address Type is not 'K'.	-2.23070	0.00000
hat_l	0	Set to 0 as HUID Address Type is not 'L'.	-2.61080	0.00000
hat_m	0	Set to 0 as HUID Address Type is not 'M'.	-2.20100	0.00000
				-2.06800

By substituting the score –2.068 (the sum of the products) as the linear estimator into the probability equation (1) presented earlier, the score is equivalent to a probability of 88.8% that the address is valid at the state level. By taking the address with the lowest score for a particular level of geography, the address most likely to be valid at that level of geography is selected. For the sake of consistency, after selecting a person's geography at a particular level such as state, when selecting the person's geography at a lower level, only addresses that agree with the previously picked address at the higher level are considered. Each lower level of geography considered is dependent on agreement at the previous level of geography. As an example:

- First, state scores are compared for each address. State with the lowest score (in the example – Indiana) is selected.
- At the next level (county), only counties within Indiana are considered. Again, the county with the lowest score is selected.

- The process is continued to the lowest level of geography (MAFID).

In some cases, the logistic regression address selection methodology will result in not assigning geography below a certain level even where it is available. For instance:

- The most likely county of residence selected for a given person is Bucks County, Pennsylvania, and
- Tract or block geocoding is not present on the Bucks County address.
- A second address for this person includes MAFID, block, and tract information in a different Pennsylvania county.
- The CPR address selected would render the geography below the county level unspecified (blank).

Such a methodology is deemed non-problematic because the belief is – better to put a person within the county they most *likely* reside rather than in another county simply because geography that is more specific is available for another address.

For persons with a choice of location, the logistic regression model improves on the StARS 2000 address selection methodology on the order of 6 – 8 percent; affecting up to more than two million persons. After developing the model, a direct comparison to the current (StARS 2000) CPR address selection methodology was conducted. Address selection using the proposed model for all persons within the 1% sample was compared with the results of known geography (from Census 2000) for the persons in the 1% sample. Comparative results are summarized in the following table:

Table A-3. Selection Results Comparison – StARS 2000 CPR and Logistic Regression Model

	State	***County***	***Tract***	***Block***	***MAFID***
Total number of persons in sample with known census geography and more than one possible location shown in administrative records (StARS Linked Person File)	92,693	193,218	315,768	352,167	381,269
Among total the number with correct location selected for CPR	68,156	137,154	213,600	231,300	246,703
Percentage of Total	73.5%	71.0%	67.6%	65.7%	64.7%
Among total the number with correct geography selected by newly developed (logistic regression) model:	75,412	148,687	235,089	254,055	268,180
Percentage of Total	81.4%	77.0%	74.4%	72.1%	70.3%
Difference between new model and CPR:	7,256	11,533	21,489	22,755	21,477
Percentage of Total	7.8%	6.0%	6.8%	6.5%	5.6%
Projected number of persons for whom state geography can be corrected using new model	725,600	1,153,300	2,148,900	2,275,500	2,147,700

In percentage terms, the improvement of the Logistic Regression Model over the current methodology is small but significant. However, in absolute terms the use of the Logistic Regression Model rather than the current methodology will correct the geography for several millions of persons within the CPR. The StARS 2008 Work Group adopted use of the LRM despite the added complexity of computation within the CPR program.