THE STATISTICAL ADMINISTRATIVE RECORDS SYSTEM: SYSTEM DESIGN, SUCCESSES, AND CHALLENGES

Dean H. Judson

Administrative Records Evaluation and Linkage Group
Planning, Research and Evaluation Division
U.S. Census Bureau

#DRAFT Current as of 11/16/00#

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion. ***Any individual level example in this paper is entirely fictitious; any resemblance to any person, living or dead, is purely coincidental****.*

Table of Contents

Acknowledgements

The construction of the Statistical Administrative Records System is a monumental effort with dozens of participants. It has taken vision, scope, many hundreds of person-hours and years of developmental time. We'd like to acknowledge the contributions of the following U.S. Census Bureau employees:

Bashir Ahmed	John Lukasiewicz
D. Mark Bauder	Esther Miller
Mikhail Batkhan	Tamany Mulder
Mike Berning	Daniella Mungo
Barry Bye	Nancy Osbourn
Ralph Cook	Arona Pistiner
Joseph Conklin	Ron Prevost
Ann Daniele	Dean Resnick
Benita Dawson	Douglas Sater
Matt Falkenstein	Doug Scheffler
Eleni Franklin	Kevin M. Shaw
Fred Holloman	Diane Simmons
David Hilnbrand	Larry Sink
Norman Kaplan	Amy Symens-Smith
Vickie Kee	Cotty Smith
Francina Kerr	Herbert Thompson
Jeong Kim	Deborah Wagner
Myoung Ouk Kim	Phyllis Walton
John Long	Signe Wetrogan
Charlene Leggieri	David Word

Abstract

The purpose of this paper is to document the scope of administrative records use at the Census Bureau both historically and currently, document the Statistical Administrative Records System (StARS), and identify successes and current challenges. The paper has four parts:

In the first section, we begin by defining administrative records and describing the current status of administrative records use at the Census Bureau. We focus in detail on five major applications associated with the StARS database, including 1) an administrative records census simulation, 2) nonresponse followup substitution and item imputation, 3) triple system estimation, 4) linking to ongoing survey programs, and 5) uses in population estimation.

In the second section, we describe the Statistical Administrative Records System (StARS), a major database project designed to support each of these applications. The StARS is a data warehouse consisting of seven major Federal databases in virtually their entirety: The IRS 1040 Master file, IRS Information Returns file, Selective Service registration file, Medicare Enrollment Database file, Indian Health Service patient file, Housing and Urban Development Tenant Rental Assistance System file, and the SSA Numident file. We describe the characteristics of these files and overview the processing they undergo to become the StARS demographic database.

In the third section, we describe some successes in the administrative records area. These include developments in modeling and calibration of administrative records data, a demographic analysis of the Census Numident/Person Characteristics File, current triple system estimation research, the AREX 2000 experiment, and successful developments in security, privacy and confidentiality agreements.

Finally, in the fourth section we describe current challenges to the use of administrative records in general. Major current challenges include: Recognizing the source of administrative data and tracking potential for error and misinterpretation at each step of the data flow; developing an understanding of the implicit database ontologies built into the source data files, and developing methods to translate from the source database ontology into an appropriate census ontology; developing methods for handling address information that is difficult to place "on the ground"; developing unduplication and record linkage methodologies; dealing with variations in data from different sources and limited microdata content in the source files; and finally, recognizing the reality that administrative records data handling is distinct from "point in time" survey data collection in that information states are constantly changing as the databases attempt to track a dynamic and constantly changing population.

Introduction: Potential Uses of Administrative Records at the U.S. Census Bureau

In the National Research Council's report, *Modernizing the U.S. Census*, (Edmonston and Schultze, 1995:167), the panel noted that administrative records data are "a major resource, both potential and realized, in the development and production of small area estimates" and further evaluated the "radical alternative" to traditional census-taking offered by an administrative records census. They recommended that research proceed on both fronts.

For the first front, recent proposals from the Census Bureau itself, including the post-2000 Estimates Conference, American Community Survey (Taeuber, Lane, and Stevens, 2000) and related work, have highlighted the continuing need for detailed, up-to-date demographic estimates for areas as small as census tracts. The current method of estimating states and counties relies heavily on administrative records data sources (Batutis, 1994; Judson, Popoff, and Batutis, 2000), and various methods for using administrative records in small area estimates are being evaluated (Weidman and Alexander, 1999).

At the same time, for the second front, the U.S. Census Bureau has been actively pursuing potential uses of administrative records databases for decennial applications. During the debates surrounding the use of sampling for nonresponse followup in Census 2000, several proposals were suggested for direct use of administrative records, ranging from direct substitution of administrative data for nonresponding households (Zanutto, 1996; Zanutto and Zaslavsky, 1996;

1997; 2001), to augmenting the Master Address File development process with U.S. Postal Service address lists (Edmonston and Schultze, 1995:103), to simulating a complete "administrative records census" itself (Myrskyla, 1991; Myrskyla, Taeuber, and Knott, 1996; Czajka, Moreno, and Shirm, 1997; Bye, 1997; Czajka, 1999).

The purpose of this paper is to document the scope of administrative records use at the Census Bureau both historically and currently, and identify successes and current challenges. The paper will have four parts: We will begin by defining administrative records and describe the current status of administrative records use at the Census Bureau, focusing in detail on five major applications associated with the Statistical Administrative Records System (StARS) database. Next, we will describe the StARS, identifying its major features. Third, we will describe some successes in the administrative records area. Finally, we will describe current challenges to the use of administrative records in the StARS database.

What are administrative records?

According to the American Statistical Association Ad Hoc Committee on Privacy and Confidentiality (1977), AAdministrative records are collected and maintained for the purposes of taking action on, or controlling actions of an individual person or other entity. The actions include such functions as licensing, registration, inspection, insuring, training, regulating, servicing, diagnosing, treating, charging, paying, or conveying other benefits or penalties. These records were not designed to count individuals, nevertheless, administrative records do provide us with count information" (see also Brackstone, 1988, and Judson and Popoff, 1998). Like

other data warehouses, initially administrative records data are obtained for non-research purposes. However, though these data are not collected with the goal of statistical analyses, in many cases we can use them in a statistical fashion, such as producing summaries, producing frequencies, and analyzing trends.

Leggieri and Prevost (1999) outlined several major projects associated with administrative records databases at the U.S. Census Bureau:

1. Administrative records should be used to improve or target improvements to the MAF/TIGER System, and to classify addresses by their commercial and residential uses.
2. Focus on research projects that will yield short-term successes demonstrating tangible benefit; build a history of evidence with iterative successes (e.g., reduced cost, response burden, program and process improvement).
3. Determine the coverage of administrative records for decennial non-respondents (and for the full universe). Determine the accuracy of content (race, ethnicity, gender, and age characteristics of the population).
4. Provide research support for expansion of small area estimates to smaller geographic units including improving input data and geographical coding.
5. Develop administrative records to the point where they can be used to support survey sample design/stratification, sample selection and screening.
6. Use administrative records to link business and demographic data and conduct longitudinal studies -- support the Longitudinal Employer - Household Dynamics (LEHD) Project and review SSEL addresses.

7. Measure the accuracy, availability, operational feasibility of utilizing administrative records to measure response error/bias and replace questions on periodic surveys that typically have high nonresponse rates (or imputation of missing data).
8. Compare administrative records with Census 2000 results. Start with aggregate comparisons, although ultimately we may have to demonstrate micro-level comparability to gain full support.
9. Consider uses of aggregate administrative records data (like food stamps, and Internal Revenue Service income data for SAIPE) for modeling and estimates (e.g., unemployment data, health insurance, school lunch program).
10. Research should support "expanded" uses of administrative records that are consistent with Census Bureau mission and data suppliers' approved data uses.
11. Conduct decennial testing in 2001-2003, to prepare the agency for a major test in 2005.
12. Research the definitional differences between demographic variables and their administrative record counterparts; determine if we can bridge these differences or if we must consider vising operationally defined demographic variables.
13. Develop appropriate messages about statistical uses of administrative records.
14. Explore how administrative records could be used to enhance American Community Survey (ACS) data.

Progress is being made on virtually all of these fronts. We will deal with five specific applications in detail.

Administrative Records Census Experiment in 2000 (AREX 2000)

An *administrative records census* (ARC) is a census whose primary source of data is administrative records. Among the important purposes of the United States Census are to provide data for the reapportionment of congressional seats, and for review of redistricting plans. In order to serve the latter purpose, the U.S. Census must produce counts by age, race, and Hispanic origin at the *block* level (Czajka, et. al., 1997). Thus, for an ARC to replace all or part of the U.S. Census, it should provide data at the block level or lower.

The National Research Council concluded that an administrative records census was not feasible for the year 2000 (Edmonston and Schultze, 1995:68). At the time, the expertise needed for performing an administrative records census was not in place in the United States, and the legal and public policy development necessary for such an idea had not yet been put into place. Instead of pursuing an ARC at that time, an ambitious research program was put into place to evaluate whether an ARC could provide the appropriate redistricting (that is, short form) data at the block level of geography.

The Administrative Records Experiment 2000 (AREX 2000) is a *simulation* of an Administrative Records Census (ARC). It is being conducted with Census 2000 for the primary purpose of evaluating the feasibility of using an administrative records census to supplement or replace the traditional U.S. Census, and to compare two methodologies for conducting an administrative records census. The model for this simulation is a report developed by Barry Bye under contract (Bye, 1997), drawing on Knott (1994), Marquis, Wetrogan and Palacios (1996), Prevost (1996),

and others. Bye=s primary contribution was in designing the operational implementation of an ARC.

In order to perform an ARC simulation, the following steps must take place: The sites for the simulation must be chosen; the administrative databases that have the highest overall population coverage must be defined and obtained; these files must be edited to standardize their concepts and, where necessary, translate into census concepts; field operations must be defined and executed; and, of course, post-processing of the composite files must be performed. With the exception of the last phase, all of these are currently in place for the AREX 2000.

Nonresponse Followup Substitution, Modeling and Item Imputation

Since the late 1980's, the Census Bureau has invested a major research effort in using administrative records data in concert with sampling for nonresponse followup (NRFU). In late January, 1999, the U.S. Supreme court ruled that sampling for nonresponse followup was not consistent with the Census Act for providing apportionment counts for legislative representation[1]. Based on this decision, the Census Bureau removed plans to sample for nonresponse followup, and reduced the sample size of the post-enumeration survey (known in 2000 as the Accuracy and Coverage Evaluation, or A.C.E.). However, research on the use of *administrative records* as a tool for handling nonresponse followup has continued. The research effort is represented by a series of papers by Zanutto (1996), Zanutto and Zaslavsky (1996; 1997; 2001), and Larsen (1999).

[1] For the record, the Supreme Court did not rule on the "constitutionality" of sampling, nor does the ruling disallow use of sampling for redistricting uses (cf. Anderson and Feinberg, 1999).

The basic goal of using administrative records for nonresponse followup is to reduce the enormous cost of the nonresponse followup operation, without a substantial loss in data quality. Leggieri and Killion (2000) note an early estimate from the 2000 census that a 5% reduction in the nonresponse followup operation could save on the order of 70 million dollars in total census costs). However, the exact method for using administrative records in this way has not been fully established. Zanutto and Zaslavsky (1996) tested direct substitution of administrative records data for nonresponding census household data, and found that direct substitution resulted in substantial errors. They contrasted the direct substitution approach with a model-based approach (which incorporated the possibility of data errors in the administrative record), and determined that the model-based approach fit the obtained data better than direct substitution.

In order to use administrative records for NRFU, the following steps must take place: An administrative records database must be constructed with sufficient coverage such that it adds to census initial mailout/mailback coverage, methods for matching these data to individual housing units on the Master Address File (MAF) must be developed and tested, and the content of the matched households must be evaluated to determine if the administrative records household closely approximates the census household.

Triple System Estimation

The U.S. Census Bureau currently uses a dual system approach to estimating under and overcount (Wolter, 1986; Hogan, 2000). As the method exists for Census 2000, a sample of

12,000 block clusters (representing roughly 300,000 housing units) in the U.S. were chosen. An *independent* address listing operation was performed on these block clusters, and an *independent* enumeration was performed. Addresses and persons enumerated by the Decennial Census responses and those enumerated by the A.C.E. are matched, and the coverage factor of the Decennial Census is then estimated. These coverage factors are estimated for each of up to 448 post-strata. Finally, to derive estimates for *every* block in the U.S., the post-stratum-specific coverage factors are applied across non-sample blocks using a synthetic method applied to the population characteristics of each block. (See Hogan, 2000, for more extensive details on the Census 2000 A.C.E. program, or Childers, 2000, for operational implementation details.)

Dual system estimation suffers from two kinds of biases: Heterogeneity in capture probabilities within post-strata, and correlation in capture probabilities across systems (known as correlation bias; Wolter, 1986). Both of these problems, when they occur, create biases in the estimated coverage factors. Unfortunately, they are hard to evaluate. Judson (2000a), using ideas developed earlier by Zaslavsky and Wolfgang (1993), and Darroch, et. al. (1993), developed a proposal to match administrative data with Decennial and A.C.E. data on a representative sample of Census 2000 cases, thus creating a *triple system estimate*. The triple system approach allows for the analysis of heterogeneity, correlation bias, and erroneous enumerations (Biemer, 2000).

In order to use administrative records for triple system estimation, the following steps, similar to the nonresponse followup steps, must take place: An administrative records database must be constructed with sufficient coverage such that it is comparable to the entire census and A.C.E.

coverage, methods for matching these data to individual housing units on the Master Address File must be developed and tested, methods for matching individuals within households must also be tested, and the content of the matched households must be evaluated to determine if the administrative records household closely approximates the census household. Finally, appropriate theory for developing and evaluating a triple system estimate, analogous to dual system estimation theory, must be developed (Judson, 2000a).

Linking to Ongoing Survey Programs

A major application of administrative records data is linking individual data from an ongoing survey operation to comparable data in an administrative records database. The purpose of such linkage is typically to "calibrate" the survey data to the administrative data; a typical example is matching income reports from survey respondents with comparable income reports from unemployment insurance wage records or tax returns (cf. Coder, 1992; Coder and Scoon-Rogers, 1995). Heretofore, the most common method of "calibration" was to compare values at an aggregate level (for example, to add up all income reports of survey respondents and compare that sum to aggregates estimated externally; see Moore, Stinson and Welniak, 2000, for a review). Such methods are used currently to adjust Current Population Survey income reports to correct for underestimation of imputed income values (Nelson, 1985; David, Little, Samhuel, and Triest, 1986).

The ability to link individual records across two databases substantially increases the analytic possibilities of both (Scheuren and Winkler, 1993; 1997). For example, in current discussions of

the definition of poverty, several commentators (cf. Short, Garner, Johnson, and Doyle, 1999) have suggested that a "net income" figure, adjusted for travel to work and child care costs, will make a better definition and remove disincentives to work. In order to evaluate this proposed measure, Sisson (2000) has proposed modeling a household's tax return status from reports provided by Survey of Income and Program Participation respondents. *In order to evaluate his tax return model*, Sisson proposes linking his *modeled* tax return status with the individual household's *actual* tax return in the IRS 1040 master file.

However, as Moore, Stinson and Welniak (2000) point out, when a survey response does not match an administrative records datum, it is too tempting to simply assume that the survey respondent is in error. "Data from independent sources are almost never completely comparable to the survey data –due to sampling frame difference, timing differences, definitional differences, etc.—and the adjustments necessary to make them comparable are often inadequate" (Moore, Stinson and Welniak, 2000:3). Currently, the phrases "calibration" or "correcting for systematic differences between the two systems", rather than "error", are more common among those researchers linking administrative with survey data (Zanutto and Zaslavsky, 2001).

In order to use administrative records for linkage to ongoing surveys, the Census Bureau has developed the capability (both technical capability and legal agreements regulating that capability) to directly link individual records from survey respondents to the corresponding individual records from some administrative data. However, when match rates vary, the potential for a "nonmatch bias" (similar, conceptually, to "nonresponse bias"; Lahiri and Larson;

2000) exists. If certain kinds of persons are more likely to fail to be matched, then those persons are not completely represented in the analysis. Research continues on matching strategies and analytic uses of matched data sets.

Population Estimation Uses

Two significant users of administrative data are the population estimation and population projection branches of the Census Bureau. Since 1993, the Census Bureau has used a components of change method referred to as the Administrative Records Method (Batutis, 1994; Judson, Popoff, and Batutis, 2000) for county total population estimates, and similar methods for Demographic Analysis evaluations of the Decennial Census (Hogan and Robinson, 1993). The key assumption underlying the Administrative Records Method is that each of the components which constitute total population change can be represented by an administrative data series. Separate administrative records series are selected to represent each component to estimate the change in population from July 1 through June 30th of the prior year. Total change, the sum of the change for each individual component, is added to the base to arrive at the population estimate as of July 1st of the current year. This method has several practical advantages over earlier methods:

- As a result of the method, several of the components of change are treated independently, which creates the opportunity for more disaggregated analysis of components;

- The method does not depend on individual states for state-specific information, and each state's and county's population is estimated in a consistent manner;
- The estimates are generated from data that are, in most cases, directly available to the Census Bureau; and
- Components represented by administrative records are more straightforward to describe to policy makers than regression-based methods.

In order to use administrative records for population estimation, the following steps must take place: Relationships must be established between the administrative records record count and the true population count within each component, and where they differ, adjustment or rake factors must be developed; appropriate geocoding mechanisms must be established for placing the administrative record into the correct geography; methods to estimate unknown or unknowable components must be developed; and finally, an assessment of biases and potential corrections for these biases must take place.

The Statistical Administrative Records System: Design

Each of the uses and potential uses of administrative records listed in the previous section require research and development, and research into operational implementation. As a central part of plans to use administrative records at the Census Bureau, Prevost (1996) proposed a "Statistical Administrative Records System" (StARS) database development project. The StARS is a research project designed to build databases of personal and address data using administrative records from various government agencies, primarily for application to decennial census research and development. It shares most features of a typical "data warehouse" (Inmon, 1996).

For output of the StARS database, we have two goals:

- For **person** data: One output record per person, assigned to an individual residence corresponding as closely as possible to Census residence definitions, in a household structure corresponding as closely as possible to Census household structure, containing microdata corresponding as closely as possible to Census short form microdata, and excluding persons which are not in the population of interest[2].
- For **address** data: One output record per individual housing unit at a Basic Street Address, geocoded to Census TIGER geography, with address microdata and concepts corresponding

[2] There are several examples of persons who are legitimately in administrative records databases yet should not be in an enumeration population. For example, proxy recipients of medicare beneficiaries , if they are enumerated on their own, should not be double counted; similarly, there are deceased persons who legitimately exist on an administrative record (e.g., their estate files a 1040), yet of course are not in an enumeration population; or, an emigrant from the U.S. might appear in an administrative records database yet of course not be a *de jure* U.S. resident.

as closely as possible to Decennial Master Address File (DMAF) address fields and concepts, and excluding locations which are not in the population of interest[3].

Source files

Seven files are used to develop the StARS database. For each file, we indicate the approximate number of individual records, either person or filing unit based, in the file. For the data content within the file, the Census Bureau requests the approximate content equivalent to "short form" data, or data that are used for other "long form" modeling (e.g., income). In any case, programmatic data not immediately useful for decennial applications are not requested. These seven files include[4]:

- IRS's Individual Master File-1040 Returns (117 million records)
- IRS's Information Returns Master File (770 million records)
- HCFA's Medicare Enrollment Database (55 million records)
- SSA's Numident File (750 million records)
- HUD's Tenant Rental Assistance Certification System (3.3 million records).
- Selective Service System's Registrant File (13 million records)
- Indian Health Service's Registration File (2.6 million records)

[3] As with persons, locations such as Post Office Boxes, commercial mailing services, or business operations are not in the housing enumeration population. Special Places and Group Quarters are an additional complication.
[4] Throughout this paper, we will use the following acronyms: IRS for the IRS 1040 file; 1099 for the IRS information returns file; HCFA for Health Care Financing Administration; SSA for the Social Security Administration; HUD for the Department of Housing and Urban Development; SSS for the Selective Service System; IHS for the Indian Health Service; USPS for United States Postal Service, "SSA Numident" for the SSN recipient database maintained by SSA, "Census Numident" for the Census Bureau version of that file.

In Fiscal Year (FY) 2001, current plans include expanding on FY2000 population and housing research by creating household statistics and exploring the following national administrative records systems for their ability to provide enhancements:

- HCFA's Medicaid recipient file
- HUD's Computerized Homes Underwriting Systems *(FHA loan application file*)
- USPS's NCOA/LACS *(National Change of Address and Local Address Conversion System*
- Education's FAFSA- Free Application for Federal Student Aid *(Student loan and grant application database*s)

-- Insert figure 1 about here --

Figure 1 provides a broad overview of the handling of these files. As can be seen in this highly stylized diagram extracted from StARS design and file management specifications, the six source files (IRS 1040, IRS information returns, HUD-TRACS, Selective Service, Indian Health Service, and Medicare) are edited to standardize name, address and other demographic information, then combined and unduplicated. An additional file, the Person Characteristics File (PCF), is derived from the SSA Numident, and is considered a "lookup" file for demographic characteristics. After the social security numbers on individual records pass through a validation algorithm, the demographic characteristics for each SSN are extracted and merged with the source files. The final result for each SSN is a "composite person record" and a "composite person file." Note that the final demographic characteristics for the composite person record may

or may not be those recorded in the PCF file. Rather than consider one data source to be uniformly better than another, the design of the StARS database makes a "best data" judgement on a field by field basis; thus, for example, a person who is in the Selective Service database receives a gender code of "male" regardless of the PCF gender. (Detailed specifications and documentation on the processing decisions that create the composite person record will be discussed later, and can be obtained from the Administrative Records Research Staff.)

Address information from the source databases has its own editing, standardization, and unduplication flow. After the composite person record is developed, address information is "reattached" to the composite person record. The result is the "person output," which feeds into post-processing and will be used for later analysis. The composite person file product is, for all intents and purposes, the StARS database.

Characteristics of the Files Included in the StARS System[5]

One feature of administrative records databases that make their use substantially different from typical survey research databases is that each database has its own unique programmatic requirements, and hence its own unique ways of categorizing data and sources of error. In this section, we will briefly describe the major features and questions about the seven files used in the StARS database.

[5] Killion, 1999, provides basic information on each file, which is used in this section extensively.

IRS Individual Master 1040 File

The most important file from the point of view of population coverage is the IRS Individual Master 1040 file. Sailer, Weber, and Yau (1993), and Czajka, Moreno, and Shirm (1997) estimate the ratio of IRS persons to population is close to 97%. Because this ratio is not based a direct person-by-person comparison, it is not an estimate of coverage per se. However, as it is close to 100%, it suggests that 1040 coverage of the U.S. population should be high. An important additional feature from the point of view of population coverage (and differential undercoverage) is that households below the filing threshold *do not need to file*. Thus, any area having a substantial level of poverty will tend to be *undercovered* by this file. Judson, Popoff, and Batutis (2000) document that this feature of the IRS 1040 file is a likely cause of biases in net migration rate estimation for the purposes of making county population estimates. A final coverage issue associated with this file is that the Census Bureau receives Social Security Numbers for up to *four* dependents only. This has not affected the estimates uses of the 1040 file, because migration estimates currently count the number of exemptions rather than enumerating individuals. However, as the estimates program attempts to develop migration estimates by demographic characteristics (see, e.g., Long and Wetrogan, 1990, or Miller, Judson and Sater, 2000), this potentially biases coverage with respect to large families[6].

In addition to coverage questions, the file itself has certain anomalies that must be accounted for. First, the file represents *Tax Year* data (which is neither a calendar year, nor a fiscal year, nor is it

[6] The author thanks Charlene Leggieri, Assistant Division Chief for Administrative Records, for pointing this out.

a point in time). Thus, April, 2000 refers to Tax Year 1999 (TY'99). We note that since extensions for filing occur, the file is processed by IRS in 52 weekly cuts throughout the year. Thus, the (almost) full TY '99 file arrives at the Census Bureau in October, 2000. Rubin (1987) noted that later filers are substantially different than earlier filers (they tend to have more complicated tax returns, as one example), thus we cannot simply take an earlier cut of the file without risking the equivalent of nonresponse bias.

It is also important to note that business entities, estates, and other institutions are included in the 1040 file. Business entities in particular must be removed. However, estates, obviously representing a person that has died, do not necessarily report *when* the filer died. Thus if one were interested in an April 1, 1999 census date, one could not determine from the data file whether the filer was alive or dead on that date. A related anomaly in the 1040 file is that a *tax filing unit is not the same as a household*. Czajka (1999) documented that 10-20% of addresses in the 1040 Master File are Post Office Boxes, business addresses, or tax preparers' addresses.

Finally, the 1040 file has limited microdata content: Prior to TY'95, the SSNs of dependents were not used by the Census Bureau, and hence were not requested or recorded for the entire file. Further, age, race, sex, hispanic origin, and date of birth microdata is not available on the file. Finally, the name information for dependents is limited to the first four characters of the name. This limits the ability of SSN validation algorithms to validate the name/SSN/date of birth on the dependent record.

IRS Information Returns (1099) File

The second major file, and the file of greatest sheer number of records in the StARS system, is the Information Returns (1099) file[7]. Prevost (1997) makes the argument for using the 1099 in addition to the 1040 file, arguing that taxpayers who do not file 1040's are still likely to be found via their information returns. Thus, the overall coverage of the combined files should be enhanced. (In a study by Huang and Kim, 2000, of their representative 1% sample of cases, 11.1% of the cases were "in addition" to IRS 1040 cases, and 4.0% of the cases came from the 1099 file.)

Like the 1040 file, the 1099 file provides tax year data; April, 2000 refers to "tax year" 1999. Similarly, the TY '99 file arrives October, 2000. Like the 1040 file, business entities, estates, and other institutions are also included in the 1099 file. The file consists of about 770 million individual records. Again like the 1040 file, a recipient address is not a housing unit, about 10-20% of the provided addresses are Post Office Boxes, business addresses, and tax preparers, and it has extremely limited microdata content.

Selective Service File

The Selective Service registration file consists of about 13 million records. Registration with the Selective Service System was required of males aged 18-25 in 1940, suspended in 1975, and resumed in 1980. Presumably, males 18-25 are required to inform SSS when they move, although we believe that there is no extensive enforcement of this regulation and hence the address information is not maintained over time. Females, non-immigrant aliens, and

[7] This file is often referred to as the "1099" file because 1099 returns make up the largest portion, about 50%. However, W-2's account for about 18% of the file as well (Prevost, 1997).

hospitalized, incarcerated, and institutionalized males, and members of the armed forces are exempt from registration. Like the 1040 and 1099 files, this file has limited microdata content. Unlike the 1040 and 1099 files, it has some content--e.g., because only males are required to register, gender is implied, and a date of birth is recorded.

Medicare Enrollment Database (Medicare):

The Medicare Enrollment Database contains current and historical Medicare enrollment cases. Since it is an historical file, it contains both "active" and "inactive" cases. The database contains 35-40 million active records at any one point in time; in September of 1993, it contained 77 million *total* records (including both active and inactive records). Coverage of the 65 and older population is believed to be high (in the range of 90-102%; Kim and Sater, 2000), but not perfect and coverage is unevenly distributed geographically. The most important current Census Bureau use of the database is in the county level population estimates system; for the population aged 65 and older, the medicare enrollment is used as an almost direct count of this population (Batutis, 1994; Judson, Popoff and Batutis, 2000).

A notable feature of the Medicare file is that it contains "proxy recipients" on the file. An example of such is John Doe receiving benefits in care of Jane Doe; or John Doe receiving benefits in care of his nursing home. Thus, when an address is listed on the file, it is often not obvious whether the address refers to the residence of the proxy recipient (e.g., Jane Doe), the address of the recipient (John Doe), or even an institutional address (e.g. the nursing home).

An additional feature of the Medicare file is that a small portion of records at any point in time are probably deceased. Kim and Sater (2000) analyzed the file at a microdata level and determined that there existed contradictory codes for a small portion of records: For example, a beneficiary would be coded as receiving benefits yet would also be identified as deceased, or, conversely, would have benefits terminated "due to death" but would not have a date of death or fact of death recorded. In the latter situation, we expect that there would be little or no incentive for the data to be corrected, since the recipient is deceased and payments are not being made. Thus, the person becomes "statistically immortal" (a phrase to describe a person who's death is never recorded in a statistical database; Abraido-Lanza, Dohrenwend, Ng-Mak, and Turner, 1999).

Indian Health Service patient file:

The Indian Health Service (IHS) patient file consists of about 10 million transaction records on recipients of Indian Health Service benefits. It has two important roles in the StARS system: First, to attempt to cover the traditionally undercovered Native American population; and second, for use in the race model to predict those of Native American race.

It is important to keep in mind that a transaction record is not the same as a person record. A person can occur with multiple transactions; and, for example, if the person provides a social security number on one occasion and not on another, it is difficult to unduplicate the transactions to create "person" records. The about 10 million patient records become about 2 million unduplicated SSNs upon completion of processing of this file.

One feature of the IHS file is that it contains many missing SSNs. About 20% of the transaction records on the source file are missing SSNs. Of those records that did not contain an SSN, for about 1/3 of them an SSN was found via a probabilistic Numident "search" process (described later in this document), which used demographic characteristics to identify persons in the Numident file with similar characteristics (and thus find an SSN).

Housing and Urban Development Tenant Rental Assistance Certification System (HUD-TRACS):

The HUD-TRACS file is based on housing subsidy payments made by Housing and Urban Development to persons in poverty. Currently, it contains about 3.3 million records on subsidy recipients. For the HUD-TRACS file, the equivalent of census short form data is available for all members of household, although race and Hispanic origin information is only available for the head of the household. In some cases, address information on the file may represent a project or landlord address--and identifying such cases is problematic. It was believed that HUD-TRACS provided an opportunity to increase coverage of those persons in poverty who would otherwise be undercovered by the IRS 1040 or IRS 1099 files.

Census Numident File:

A most important source file for demographic characteristics is the Census Numident file. The Census Numident file is derived from the Social Security Administration's Numident file, intended to cover every transaction on every social security number issued since 1940. Like the

Indian Health Service file, these transaction records were converted into SSN-based records (one record per SSN, one SSN per record). Thus, the approximately 750 million transaction records on the SSA Numident file became about 400 million individual SSN records on the Census Numident file. For each SSN, the Numident file contains information on date of birth, gender, race, and place of birth. It is thus the most complete source file for demographic characteristics. About 35% of SSNs on file have alternate names (due to marriage, divorce, and various other name change situations; Scheffler, 1999).

Unfortunately, the Numident file contains no information on residence geography. Thus, by itself it is not suitable for placing persons in households or in any geography lower than the U.S. level. (It is possible, of course, to link it with files that do have geographic information; this is its use in the StARS system.) Further, Taxpayer Identification Numbers (TINs) are not on the file; thus, a person with a TIN cannot be found in the Numident. Likewise, a new SSN issued since the last database update will not be found.

A notable feature of the Census Numident file is that about 60 million persons on the file are deceased but not identified as such (Falkenstein, Resnick, and Judson, 2000). Many of these are people who are "statistically immortal"; that is, they have almost certainly died but that fact has never come to the attention of the SSA. In all likelihood, some of them, for example, international emigrants, will never come to the attention of the SSA (Abraido-Lanza, et. al., 1999). Of course, those persons in that state have substantially different demographic characteristics than others: They are older; because of demographic shifts that have occurred in

the U.S. in the past century, they are more likely to be white, and because of differential mortality they are more likely to be female (Miller, Judson and Sater, 2000). In order to correct for this fact, Falkenstein, Resnick and Judson (2000) developed a probabilistic mortality model to predict the conditional probability that a person is deceased, conditional on the fact that they have not be recorded deceased, their last known interaction with the SSA system, their gender and their age. (For the record, we note that about 6% of the SSNs on the Census Numident are missing a gender code.)

In the Social Security Numident, race coding has changed over time. Prior to 1980, SSA recorded three races: White, Black, and Other or unknown); beginning in 1980, the race codes reflect five races (White, Black, Asian or Pacific Islander, Hispanic, American Indian or Eskimo), and included the codes Other, Blank, and Unknown. Note that one cannot assume that because these codes exist on the file, they are equally representative or likely; for example, 20% of the records in the Census Numident are either Unknown or Other, reflecting the fact that information states have changed over time[8]. In addition, about 25% of the SSNs in the Census Numident have transactions with different race codes. This is handled by a decision rule that determines an "estimated race"[9] from the information available in all transactions associated with that SSN (U.S. Census Bureau, Census Numident Programming Specification, 1999).

[8] We will explore this feature of administrative records, that information states change over time, extensively later in this paper.

[9] In the Numident programming documentation, the term "best race" was used. This term has been dropped from use.

After 1985, the Social Security Administration instituted enumeration at birth—a program designed to enroll infants and provide them with social security numbers. This corresponded approximately with changes in the tax law that required households to provide SSNs for child exemptions they were claiming, if the child was two years old or older. Unfortunately, enumeration at birth does not include a race classification. Thus, *over time*, our ability to determine a person's Census-defined race from the race coding on the Numident will likely deteriorate (Miller, Judson, and Sater, 2000, document that this is probably already occurring).

Person and Address Editing

The first phase of data handling of a source file is the assignment of an unique identifier (UID) to each record in sequence. These UID's serve as pointers allowing navigation through the various files—for example, if a particular record in the composite person record appears to have erroneous data, then we can use the UID to navigate *all the way* back to *all* of the source data that led to the construction of that record. Immediately thereafter, addresses are split from persons, and each are handled in their own way, with pointers (linkages) to rejoin them in subsequent processes.

Obviously, each source file records data in its own fashion. For example, IRS 1040 names are recorded as a 35 character string, with a 35 character supplemental field; in the case of a "married filing jointly" return, *both* names will appear in the name string; thus, they must be parsed and placed in separate individual name fields (U.S. Census Bureau, StARS Person Edit Programming

Specifications, 2000). At this phase, in every source file, both person data and address data are parsed, standardized, and placed in common fields. The result of this handling is the "Edited Person File" (EPF) with each record being designated an "Edited Person Record" (EPR).

A similar process occurs with addresses. However, since addresses occur with substantially more variation than person names and characteristics, they required a substantially more elaborate editing process. Addresses are first passed through an address standardizer, which attempts to match the address to a database contained in CODE-1 commercial software. Direct probabilistic matching between databases was also attempted in this phase. After matching, addresses were parsed into separate address components (e.g., street name prefix, street name directional indicator, house number, street name, street name suffix, street type, within structure identifier, etc.) These components are flexible enough to incorporate different types of addresses (e.g., rural route versus PO Box versus standard house number/street name addresses). The final result of this handling is the Master Housing File (MHF). Table 1 describes the fields in the EPF and, similarly, those obtained in the MHF.

-- Insert Table 1 About Here --

Social Security Number Validation and Search

Substantial numbers of social security numbers (SSNs) are recorded in error in administrative records databases, and the number and kind of error varies by database. In a study by the Department of Health and Human Services (1990), as many as one in ten SSNs are recorded in

error in their sample of databases. This error rate varied by type of organization, with higher discrepancies in prisons and financial institutions, and the lowest discrepancies in tax collecting organizations. (Judson and Popoff, 1998, provide a table of causes of disrepancies and discussion.) Prisons, whose clients we presume tend to want to avoid detection, exhibit a median percent discrepant SSN=s of 28%. In contrast, even vital statistics bureaus, for whom we would expect high motivation to report accurately, exhibit median percent discrepant SSN=s of 11%, or close to one out of every ten SSN=s in their file. One organization, an otherwise-unidentified financial institution, had 53% discrepant SSN=s. Happily, tax collecting organizations had the lowest percentage discrepant SSN=s at 4%, probably, again, reflecting strong motivation (for both parties) to report and maintain this particular datum accurately. The Department of Health and Human Services report (1990:12) goes so far as to state: AAlthough most sample organizations consider SSN accuracy important, this concern does not appear to affect their SSN discrepancy rates@--and recall that this is a relatively *simple* data element, that *should* be easy to record. Likewise, it is of fairly high legal importance, so both respondents and record keepers *should* be highly motivated to record it correctly. Yet, these disrepancies still occur. The report indicates that name changes due to marriage appear to be a major source of discrepancies; while this is a very plausible explanation for the field most likely to cause discrepancies, it is not a plausible explanation for the next most likely cause of discrepancies, the fact that the SSN is not in the Social Security Administration file (which caused 20% of the discrepancies) and date of birth mismatches (which also caused 20% of the discrepancies).

For the StARS database, Czajka, 1999 cites a 1987 special study that found that in the IRS 1040 file, .5% of primary filer, 1.6% of secondary filer, and 3.4% of dependents' SSNs are in error. However, much of the definition of "error" depends upon how aggressively the validation algorithm attempts to correct for name, date of birth, and SSN errors (for example, name and SSN errors each suffer from transposition), while cultural practices for naming conventions vary dramatically across cultures. In particular, Asian and Hispanic cultures have distinctly non-Anglo naming conventions; Blumberg and Goerman (2000), identify a Latino example, noting that in Hispanic culture every individual has two surnames, a father's surname *and* a mother's surname. This means that father, mother, and child often have different surnames from each other.

The SSN validation algorithm in the StARS system constructs many different combinations of the first name, middle name, and last name of the source record, and additionally attempts to catch simple SSN transpositions, particularly in the digits near the end of the SSN. For those name/SSN/date of birth combinations that do not validate, a search algorithm attempts to use probabilistic record linkage techniques based on the Fellegi-Sunter model of record linkage (Fellegi and Sunter, 1969; Copas and Hilton, 1990; Winkler, 1995). Probabilistic linkage in this case is based on name, age, gender, and other characteristics of the person; if no corresponding demographic characteristics can be found in the Census Numident file, the record is declared not validated and not found in search. (See U.S. Census Bureau, Social Security Number Verification Programming Specification: Social Security Number Verification Against the Census Numident, 2000.)

Development of the Person Characteristics File (PCF)

The portion of StARS dealing with statistically generated demographic data is the Census Person Characteristics File (PCF). The Person Characteristics File differs from the other files in the StARS database in that it does *not* contain current address information, but *does* contain a reasonably detailed profile of a person's demographic characteristics. It was generated from the Social Security Administration's (SSA) Numident file, an approximately 750 million record file that has a record of every transaction associated with every Social Security Number (SSN) ever issued[10]. Also included is other micro level information for SSN holders. In the autumn of 1999, the Census Bureau obtained a copy of the entire SSA Numident file (with the exception of data fields SSA was prevented from sharing by prior agreement or statute). The SSA Numident file was edited and unduplicated for use by Administrative Records Research (ARR) in the Planning, Research and Evaluation Division (PRED) of the Census Bureau (U.S. Census Bureau, Numident Programming Specification, 1999). The edited and unduplicated version of the Numident file used by ARR, called the "Census Numident," will be referred to exclusively from here on.

The completed PCF is essentially a subset of the Census Numident file, with additional mathematically calculated variables provided. Four modules transform the data from the Census Numident to become the PCF:

[10] Recall that a transaction is not a person; strictly speaking, an SSN is also not a person. Thus, each SSN may have multiple transactions (application, name change, other interactions) associated with it, and each person can

- The Race Module, developed by Bye (1998) and Bye and Thompson (1999), estimates the race and hispanicity of the SSN holder. It was needed because the Numident did not categorize race according to Census definitions.

- The Gender Module, developed by Thompson (1999), estimates the gender of SSN records that do not have a gender recorded (about 6% of the Census Numident).

- The Mortality Module, developed by Falkenstein, Resnick, and Judson (2000), estimates the probability that a person in the Census Numident file is deceased. (Recall that about 100 million records exist in the Census Numident file that should be recorded as deceased but are *not* so recorded; see Falkenstein, et al, 2000.)

- The Random Number Module generates a collection of pseudo-random numbers, which are stored with the PCF and used in the race, gender and mortality modules.

Person and Address Processing

After the multiple files are edited, SSNs are validated (and invalid name/SSN pairs are corrected by the search process), and the person characteristics file is constructed to provide a lookup for demographic characteristics, it is time to combine information. As noted above, combining often-inconsistent information across databases is a challenging task. It requires a complex set of decision rules when information differs. In effect, when the source files differ in content, the decision rule decides what a person "really" is.

---

conceivably have none, one or several SSNs associated with him or herself. Note that an SSN can be held by a

As an example of such a decision, let us consider the race coding decision rule, reproduced below. (Note that race coding decisions only need to be implemented when there are multiple races reported in different files.) The decision as to which race to assign to a person occurs using the following rules:

a) If the IHS record reflects American Indian/Alaska Native, retain that race.

b) Otherwise, select the most frequent observation.

c) If there any ties, the PCF acts as a tiebreaker.

d) Numident values are weighted only once per SSN (to preclude favoring the Numident value).

A test of the selection rules against the entire first cut of SSN verified records yielded numbers reasonably close to the U.S. population breakout, according to national estimates. Test numbers (reported in unpublished technical memoranda) are as follows: Whites represent 82.2% of the test cut, Blacks 12.8%, Asian/Pacific Islanders 4.0%, and American Indian/Alaskan Natives 0.9%.

Similar selection rules have been developed for dates of birth and death, gender, Hispanic origin, and other demographic characteristics. Selection rules are being developed for addresses, although addresses have a much more complicated structure.

The Final Composite Person File

---

deceased person; thus, an SSN may correspond to no living person, as well.

Construction of the final composite person file is currently waiting on the development of final address selection rules at the U.S. Census Bureau. A draft of the composite person record layout is given as table 2.

-- Insert table 2 about here –

Administrative Records Successes to Date

Modeling and calibration

To date, much research in administrative records has focused on modeling data from one database to another, and hence "calibrating" one to another. As an example of such research, the race model (described in Bye, 1998 and Bye and Thompson, 1999) is an attempt to deal with both the 20% of the Census Numident that has no race information, and to convert the race coding from the SSA codes into four races and two ethnicity codes. The race model attempts to impute race and Hispanic origin based on Social Security Administration (SSA) Numident data combined with data from the Indian Health Service (IHS), Spanish, and Asian name lists, and geographic racial and ethnic distributions[11].

The accuracy and completeness of race information on the SSA Numident are affected by many factors: 1) In the initial stages of the Social Security Program, SSNs were issued on a SS3 Internal Revenue form, which did not ask for race information; 2) The reporting of race on the SSA applications has always been voluntary; 3) Race categories have changed over time; and 4) Original application data are not available for people who filed claims for Social Security

[11] For the record, two race models were estimated: The first did not incorporate geographical information (called the Aone-level@ model; Bye, 1998); the second incorporated geographical information in a multilevel model (called the Atwo-level@ model; Bye and Thompson, 1999). When attempting to implement that Atwo-level@ model, the required geographical location information was not available on the Census Numident and could not be constructed properly. Thus, for the 1999 Person Characteristics File, only the Aone-level@ model is used while research into implementation of the Atwo-level@ model continues.

benefits prior to 1970, the year when the electronic Numident was completed (Bye and Thompson, 1999).

The models were estimated from a data set consisting of a direct linkage of Numident, IHS, and name list data with Current Population Survey (CPS) race and Hispanic origin responses for the 1991 and 1994 March CPS. Thus, the model seeks to recreate or predict CPS race and ethnic responses from Numident, IHS, and name list information. The model=s output is a vector of probabilities that are estimates of the joint probabilities of Hispanic origin response (Yes/No) and the traditional four race responses -White, Black, Asian/Pacific Islander (AS/PI), American Indian/Alaskan Native (AI/AN).

A cross-validation report of the predicted equations by state, metropolitan statistical area, sex, age and marital status was developed by using the rotation group structure of the CPS. This was accomplished by omitting a rotation group, estimating the model on the remaining groups, and then comparing the predicted response with the actual response on the CPS omitted group. This was done in rotation for each of the eight groups and the resulting predictions accumulated. This resulted in a full set of CPS data in which the predicted responses were tabulated only for cases that were not used to estimate the regression coefficients on which the predicted response probabilities were based. (Results of this cross-validation are in Appendix A of Bye and Thompson, 1999). Although some lack of fit exists, predictions from the models are much in accord with the omitted rotation group=s reported characteristics, and, overall race distributions

are also well approximated. The difference between geographic areas actual counts and model prediction for specific race and Hispanic origin were negatively correlated with the proportion of the group in the area. In applications in which current residence is known, incorporating the local area race and ethnic distributions into the imputation process would be useful.

To discriminate between the four traditional race (excluding the race of AOther@) responses on the CPS, 24 equations were developed for race and 6 equations for predicting Hispanic origin. The 24 equations for race were made up of three equations for Hispanics and three for each of the seven possible values of best Census Numident race for non-Hispanics.

The models are fixed effects logistic regression equations, with the general structure depicted below:

$$\ln\left(\frac{p}{1-p}\right) = f(\text{namecharacteristics}, \text{Numidentindicators})$$

where,

p=P[Race=i and Hispanic=j];

Aname characteristics@ include a lookup of the person=s name against the hispanic name list; the american indican name list, and the asian name list; and

ANumident indicators@ include: How a person was coded in the two race coding schemes, and whether Hispanicity was recorded at any point in the SSA Numident.

Specific key predictor variables incorporated into most models include:

- Last Numident race;
- Variables representing the presence of a difference in race codes among multiple Numident records;
- Race as it was originally reported on the initial application for a SSN;
- Place of birth;
- Indicators of matches to the Census Bureau's 1990 Spanish Surname File and to an AS/PI surname list;
- Percentages of AI/AN in the PES; and
- An indicator from the Indian Health Service.

Note that race and Hispanic origin are simultaneously predicted in this model, and that while geography is not used as a variable in the race model for the 2000 StARS database, it is being considered for inclusion in future versions. For more model details, including specific equations and parameter estimates, see Bye and Thompson, 1999.

A second successful example of modeling and calibration is the mortality model. The mortality model is somewhat different than the item imputation models of race and gender. Instead of imputing particular data elements in the file, the purpose of the mortality model is to determine whether a particular record should be *alive in* the file. That is, the mortality model determines whether a record has in fact Aexited@ the population via death. It is now known that significant inconsistencies can exist in administrative record files. For example, the Medicare file was examined for death data inconsistencies and some problems with inappropriate death indicator flags and dates of death were found (Kim and Sater, 1999).

The mortality model treats the Numident file as the universe of interest, and applies the theorem of total probability to capture the probability that a particular SSN holder is dead or alive. In the theorem of total probability:

$$P(D) = P(D \mid R)P(R) + P(D \mid \sim R)P(\sim R)$$

where,

D is the event "the person is deceased in the population";

R is the event "the person is reported as deceased in one of our databases";

P(D) = Cumulative probability that individual is deceased (calculated from a cohort life table using the last known alive date as condition);

P(R) = Probability individual recorded as deceased on one of the six source files (calculated from the file itself); and

P(D|R) = Probability that an individual is deceased given that a source file records that they are deceased (assumed=1).

The term of interest is P(D|~R): The probability that an individual is deceased given that *no source* file *records that they are deceased*, that is, these are the people that should be flagged as deceased somewhere, but are not.

The files used for the Adeath indicator@ are the IRS 1040 and Medicare files. If the SSN holder is recorded as deceased on either IRS 1040 or Medicare files, we record their probability as 1.0.

If neither of these files records that the person is deceased, we calculate P(D|~R) from this model. Again, extensive details can be found in Falkenstein, Resnick, and Judson, 2000.

As a third example of such modeling and calibration research, Zanutto and Zaslavsky (Forthcoming in 2001) have developed methods for estimating the population of blocks using indicator data from mailback census responses, nonresponse followup and administrative data. Data take the form of a cross-classification table, and the purpose of the model is to estimate the total population size. The model itself is a so-called "loglinear" model of the cross-classification table.

PCF/Census Numident Evaluation

As noted above, the Person Characteristics File (PCF) was developed by modeling data from the Census Numident. Miller, Judson, and Sater (2000) evaluated the success of the PCF modeling process, by comparing the age, race, sex, and Hispanic origin distributions in the 1998 PCF with modeled mortality with the comparable distributions in the 1998 national population estimates. Figure 2 compares the overall race distributions for these two sources. In this figure, "White" refers to those whom the race model determined to be white, "black" refers to those whom the race model determined to be black, "API" refers to those whom the race model determined to be Asian or pacific islander, and "AI" refers to those whom the race model determined to be American Indian. After including mortality, the PCF file contained about 8.5% more living persons than the 1998 estimates; therefore, "PCF Controlled" refers to a comparison in which PCF total population is raked to make PCF total population match 1998 estimates.

-- Insert figure 2 about here --

As can be seen, the overall distribution is very comparable across the two sources, and this is enhanced when the PCF total population is controlled to the 1998 national estimates.

Figure 3 breaks out the data by Hispanic origin. As can be seen, we begin to see patterns in which the PCF estimates diverge from the 1998 national estimates. For example, the model appears to undererestimate white non-Hispanic (generating a percentage white non hispanic that is 3.2 percentage points lower than the national estimate) and overestimate white Hispanic (generating a percentage white hispanic that is 2.5 percentage points higher than the national estimate). In other data presented in Miller, Judson and Sater (2000), some individual race/Hispanic combinations (for example, API/Hispanic) are estimated particularly poorly.

Finally, to illustrate the problem of Enumeration at Birth in the Numident (and declining race codes), we present figure 4.

-- Insert Figures 3 and 4 about here --

Figure 4 makes clear that race reporting is deteriorating over time. While almost certainly some of those children in the youngest age categories will eventually report their race to the SSA, we expect that as that cohort ages, there will be less and less useable race data on the Census Numident.

Evaluations of Administrative Records Coverage

Judson, Popoff, and Batutis (1996) developed a model of biases in administrative records, and an aggregate correction factor for handling these biases. However, recognizing that the aggregate correction factor was no substitute for a microdata evaluation, they also proposed that direct evaluations of the coverage of administrative records databases take place.

Several of these evaluations have subsequently taken place. Kim and Sater (2000) analyzed the characteristics of the Medicare enrollment database, comparing 1990 Medicare counts to Decennial census counts of those 65 and older, and documented that the ratio of the two range from as low as 90%, representing undercoverage, to as high as 102%, representing overcoverage. In particular, it appeared from their evaluation that states with a substantial "snowbird" population (California, Nevada, Arizona, Florida) tended to have undercoverage, while others tended to have overcoverage.

Similarly, Pearson and Sater (2000) have begun the process of evaluating the use of individual level data from IRS 1040 returns as a way of estimating net migration rates from county to county. They have noted that filers and spouses often migrate in different directions, and that these patterns are not uniformly distributed over different states and counties. The goal of their research agenda is to develop subnational estimates of migration by full Age/Sex/Race/Hispanic origin categories.

Finally, as noted above, combining information from multiple, occasionally-inconsistent data sources is a nontrivial task. Huang and Kim (2000) analyzed a 1% sample of the multiple databases with the direct purpose of assessing overlap and field inconsistences. The one-percent sample selected from each file was drawn using the selected Social Security Numbers (SSNs) in the same way that previous one-percent samples were drawn from the IRS 1040 file at the Census Bureau. The six samples were unduplicated and merged by SSNs to form a one-percent

sample database (excluding deceased persons and those with invalid addresses) having information such as name, address and demographic characteristics (age, sex, race). Based on this sample, their findings included:

- Out of a total of 2,359,460 persons in the one-percent sample database, the number of persons from the IRS 1040 file was 2,097,213 (88.9 percent). Of the 2,097,213, 49.1 percent were primary tax filers, 20.1 percent were secondary tax filers, and 30.8 percent were listed as dependents.
- The total number of persons found on files other than the IRS 1040 in the sample database was 262,247 (11.1 percent). Of this, 4.0 percent were from the IRS 1099 file only, 2.7 percent were from the Medicare file only, and 3.2 percent were from the overlapped files of IRS 1099 and Medicare files. The percentage of persons found on files other than IRS 1040 was 9.9 percent from a combination of IRS1099 and Medicare, and 1.2 percent from files or overlapped files other than IRS 1099 and Medicare.
- Out of a total of 2,359,460 persons in the database, 49.2 percent were from one file only, and 50 .8 percent were from multiple files.
- Among a total of 1,198,597 SSNs found on multiple files, 68.4 percent of the SSNs had matched addresses; 57.9 percent had matched full name, including first, middle, and last names; 11.5 percent had the matched full name without the middle name or middle initial. Further matching based on the first and last names, when one record has the middle name but the other not, added an additional 32.1 percent of the name match to the total name match rate. The total name match amounted to 90 percent of those SSNs found on multiple files. 61.9 percent had *both* matched addresses and names.

- Among a total of 379,373 person records with different addresses, 94.5 percent had two different addresses, 86.1 percent were within the same state, and 61.7 percent were within the same 3-digit ZIP Codes. Further follow-up is needed in order to find out whether persons with multiple addresses are movers or those having multiple residences.
- In five of the six files in the database, more than 83 percent of the addresses were house number/street name style addresses. The percentage of rural style addresses in all six files was less than 11 percent and those records in the IHS file had the highest percentage of rural style addresses (10.5 percent). The percentage of P.O. Box addresses was 45 percent for the IHS file, and was less than 9.5 percent for the other five files.
- The distribution of demographic characteristics reflects the nature of each file. SSS was composed of 100% young (13-30) males. IHS was comprised largely of American Indians (84.7 percent). 84.6 percent of the persons in the Medicare file were age 65+ and, because of females' higher longevity, there were more females (56.7 percent) than males in the file. As minority female-headed households are more often economically disadvantaged, 65.4 percent of the persons in TRACS were female and 33 percent of them were Black.
- A person who appeared outside the IRS 1040 file was termed a "value added" person (because such persons presumably represent an unique contribution of that particular file toward complete population coverage). In comparison with the general population, there was a consistent shift in age distribution toward the younger age (<65) among the single source "value-added" persons for all files except SSS. This phenomenon was particularly evident in IRS 1099 and TRACS. There were more Blacks among the value-added persons (persons on files other than IRS 1040).

Triple System Estimation Research

As noted above, following early work by Zaslavsky and Wolfgang (1993), the Census Bureau has been testing triple system estimation notions. As of this point, most development has been theoretical; so this section will focus on background theory.

Traditional dual system estimation forms a 2 x 2 table, representing capture in the Census or the A.C.E. If we add a third data source, the StARS system, to the Census and the A.C.E., we are in a position to study our ability to make a triple system estimate as a way to further improve census results.

A triple system estimate, and, equivalently, an evaluation of coverage issues across databases, can be viewed as a 2 x 2 x 2 cross classification table, where the first margin indicates whether a person was captured in the decennial census, the second margin indicates whether a person was captured in the A.C.E. survey, and the third margin indicates whether a person was captured in StARS. The table appears as follows:

				StARS	
				Captured	Not captured
Census	Captured	**A.C.E.**	Captured	$Y_{1,1,1}$	$Y_{1,1,2}$

Not captured	$Y_{1,2,1}$	$Y_{1,2,2}$

			Captured	Not captured
Not captured	**A.C.E.**	Captured	$Y_{2,1,1}$	$Y_{2,1,2}$
		Not captured	$Y_{2,2,1}$	$Y_{2,2,2}$

As can be readily seen, each cell in this 2 x 2 x 2 table represents a different cross-matching outcome, and the most appropriate model for such a table is known variously as the "log-linear" model, "poisson regression" model, or, for fixed total sample size, the "multinomial" model (Bishop, Fienberg, and Holland, 1975). Using this model, various kinds of nonindependence between the 3 systems can be modeled and examined for goodness of fit. For example, we can presume that persons are captured by the three systems independently but, if they are captured by any, have an additional chance of being captured by all three (that is, there should be more people falling into the upper leftmost cell $Y_{1,1,1,}$ denoted captured in census, captured in A.C.E., and captured in StARS, than would be expected if the systems were independent). To specify this model and estimate the expected cell frequencies, we merely specify the appropriate loglinear design matrix corresponding to this model.

By comparing one model against another, we can assess the degree to which our actual cell counts correspond to those expected under either model[12]. The most crucial tests, then, are those comparing one proposed model to another (this is discussed in great detail in Bishop, Fienberg,

[12]We note that we cannot fit the cell that is missing, i.e., $Y_{2,2,2}$, and we cannot fit any margin (internal or external) that contains a zero count.

and Holland, 1975, or more recently, Judson, 1992, or Agresti, 1996). In this context, it does not appear necessary to limit our consideration to hierarchical loglinear models, i.e. those that fit all lower-order interaction terms with the inclusion of any higher-order interaction. Once cases are placed in cells in this table, then either 1) the table may be modeled directly, 2) poststrata can be developed to approximate homogeneity within poststrata (Wolter, 1986), or 3) their characteristics can be used as variables in modeling their likelihood of being placed in any particular cell. Once the modeling strategy is determined, then this model is used to estimate the (2,2,2) cell and to develop coverage factors that predict undercount for adjustment purposes. At this point, three distinct approaches to incorporating a third system have been developed.

Zaslavsky-Wolfgang approach

The first approach for incorporating the third system is represented by the work of Zaslavsky and Wolfgang (1993), in which constraints on the system are imposed and, using those constraints, estimates of the unobserved Y_{222} cell are obtained.

Space does not permit a full exposition of the methods by which the Zaslavsky-Wolfgang approach can be translated into the "loglinear" approach. Further, while we believe it to be the case that all the models considered in their paper can be so translated, we have not yet proved that conjecture, although contracts are established to verify or refute this conjecture.

Darroch, et. al., approach

In the context of dual system estimation, Mulry, Davis, and Hill (1997) argue forcefully for including an analysis of heterogeneity (in postrata[13] associated with Census/A.C.E. matches), by performing individual-level logistic regression modeling and developing predicted noninclusion probabilities. These predicted noninclusion probabilities are then used in Horvitz-Thompson style estimation methods in place of the homogeneous postrata inclusion probabilities (Wolter, 1986:340). In this study, the availability of unambiguous classification of each case provides the potential for performing the same analysis. Indeed, these analyses can be performed in any direction (i.e., using StARS data to predict classification status, as if StARS were the only data source; or using Census data to predict classification status, as if only Census data existed, etc.)

An example of the heterogeneity models is found in Darroch, et. al., 1993, in which individual-specific inclusion probabilities are estimated using a variant of the Rasch models developed for psychometric testing (Rasch, 1980). In our context, to fit a model of this kind, we merely translate our "loglinear" model into a "log-rate" model.

Biemer approach

A final approach, proposed by Biemer (2000), is the use of Latent Class Analysis to model respondent captures with respect to the 3 systems. In the latent class model, we assume that the three systems represent measures attempting to predict the latent class membership of each individual, with the two latent classes being "true census day resident" and "true census day

[13]Strictly speaking, these are not poststrata since they do not control to estimated population totals; a recent term to describe these groups is "estimation domains." Given the common practice of calling these groups postratum/poststrata, we will continue to use this terminology here.

nonresident." However, whether any particular case falls into either of these two categories is *unobserved*.

As before, we will attempt to outline a translation of this approach into the general "loglinear" model we have described. For the sake of exposition, we will presume that each case falls into one of two latent classes, "true resident" (represented by the variable T=1), and "true nonresident" (represented by the variable T=0). Obviously, we now have a mixture distribution, with cases in class T=1 following a poisson distribution with one set of parameters, and cases in class T=0 following a poisson distribution with another set. Thus, the log-likelihood function is expanded into a mixture distribution. Because this is a latent variable problem, it is particularly amenable to missing-data estimation approaches such as the E-M algorithm (Dempster, Laird, and Rubin, 1977; McLachlan and Krishnan, 1997). Upon convergence, the estimates of the parameters of the mixture distribution model are used to generate expected cell frequencies and these are in turn used to estimate the missing $Y_{2,2,2}$ cell.

Implementation and operational requirements

In order to implement this model, we must be able to unambiguously place each case, whether a person or a housing unit, in one and only one cell. In order to do so, at least two fundamental procedures must be properly developed:

1. Individual person records must be matched across the three databases; while record matching in Census/A.C.E. has been well-established, comparable record matching to administrative lists has proven to be a significant challenge (e.g. Childers and Hogan, 1984). The problem of placement of persons in specific cells is hampered by difficulties matching administrative

records addresses (often with Post Office Boxes or other non-city-style addresses) with the existing Census/A.C.E. address lists, and

2. Procedures must be developed for handling imputations, erroneous enumerations, and insufficient information cases in the administrative records database–similar to those developed for Census/A.C.E. methods in which a person is "data defined" (Zaslavsky and Wolfgang, 1993; Biemer, 2000; Childers, 2000).

It is these challenges that have hampered previous evaluations of administrative records data versus other systems (Zaslavsky and Wolfgang, 1993; Marquis, Wetrogan, and Palacios, 1996; Sweet, 1997), and, indeed, an evaluation of the Census/A.C.E Dual System Estimate itself (Ericksen, Kadane, and Tukey, 1989; Bell, 1992; Chao and Tsay, 1998; Brown, Eaton, Freedman, Klein, Olshen, Wachter, Wells, and Ylvisaker, 1999). However, given the StARS database, such a test is in the planning stages (Judson, 1999).

The AREX 2000 Experiment

AREX 2000 (see Pistiner, 1999) is an experiment conducted during Census 2000. The motivation for conducting this experiment in Census 2000 was to enable an evaluation based on comparisons of administrative records results with Census 2000 results in order to provide critical measures and a basis for 2010 Census planning.

The principal objective of the AREX 2000 is to compare two methodologies, originally proposed by Knott (1994), for conducting an administrative records census to Census 2000 and to evaluate the results and costs. Other objectives included:

- Collect relevant information, available only in 2000, to support ongoing research and planning for administrative records use in the 2010 Census.
- Compare an administrative records census to other potential 2010 methodologies.

This information, together with cost data, will provide assistance in planning major components of future decennial censuses, particularly a census that has administrative records as its primary source of data.

Knott refers to the two methods as the "top down" and the "bottom up" approaches.

Method 1: **Top-Down Approach -** A number of national-level administrative records files are assembled, unduplicated using the Social Security Numbers (SSNs), assigned block level geographic codes using the MAF/TIGER system and then the results are tabulated at the block level of geography. *Assigning records to individual housing units is not attempted, and the Master Address File is not used in this method*. This method does not provide a census of households or housing units.

This method, at a minimum, is designed to meet the data requirements for apportionment and redistricting by providing counts of the voting age (18+) population by race, Hispanic origin for blocks and counts of the population under age 18 for states.

Method 2: **Bottom-up Approach** - A number of administrative records files are unduplicated using SSNs and addresses are then *matched to a complete list of possible residential addresses on the Master Address File*. The inconsistencies in addresses are resolved (by computer, clerical review, and field verification), and, after the addresses are completely geocoded, then the results are produced.

This method meets requirements for apportionment and redistricting and provides additional 100% data (such as household relationship and tenure) and sample data (such as income).

Two sites were selected with a total of approximately one million housing units and a population of three million persons. One site included Baltimore City and County, Maryland. The other site included Douglas, El Paso, and Jefferson Counties, Colorado. The sites provide a mix of difficulty in conducting an administrative records census (See Pistiner, 1999).

Approximately one half of the test housing units were selected based on criteria assumed to be easy-to-enumerate in an administrative records census and the other half were selected based on criteria assumed to be hard-to-enumerate. Baltimore County and Douglas and Jefferson Counties are the easy-to-enumerate areas, while Baltimore City and El Paso County are the hard-to-enumerate areas. Easy-to-enumerate and hard-to-enumerate criteria are listed below.

Easy-to-enumerate criteria for a selection of test sites:

- more city style addresses with house numbers and street names
- more single-family housing units

- older age cohorts (65+)
- a non-mobile population
- white and black population

Hard-to-enumerate criteria for a selection of test sites:

- more non-city style addresses
- more multi-unit housing (i.e., rentals)
- younger age cohorts (children below 18 years of age, believed undercovered by administrative records databases)
- mobile populations as indicated by more mobile homes, immigrants, and seasonal movers
- non-black and non-white populations

Table 3 describes the 1990 distributions of race and ethnic characteristics for the five chosen sites and provides additional background information.

-- Insert table 3 about here --

Because of experiences with state and local files in the 1995 and 1996 tests of administrative records, it was determined that the state, local and private files would be very difficult to obtain, difficult to handle, and of uncertain quality (see Neugebauer, Perkins, and Whitford [1996], and Leggieri and Killion [2000] for discussion and evaluations). Thus, the focus for this experiment was on the national files of the StARS database.

Record linkage research

As noted, administrative records data represent an unique challenge for unduplication and record linkage. Obviously, these unduplication and record linkage are necessary (albeit not *sufficient*)

conditions for the proper development of administrative databases. In parallel with developments in data handling, Winkler (2000) and Judson (2000b;2000c) have worked on developing theory to improve the possibilities for record linkage. While data handling, parsing, and standardization are themselves crucial components for improving our capabilities, it is also necessary that we improve our ability to *discriminate* between matching and nonmatching data, often in conditions with substantial error and uncertainty. The current research focuses on enhancements in the EM algorithm and analogies to text classification (Winkler, 2000), enhancements in "fast lookup" methods allowing long-range searches, and taking advantage of lattice-theory in further specifying and elaborating the information content of record linkage data (Judson, 2000b;2000c).

SSN Validation and Search

When attempting to unduplicate a file, or link records across files, particularly when the files are of substantial size, it is very important to minimize the unduplication or linkage work. One way to do this is to make sure that the file(s) contain unique identifiers. In 1999, in the context of an evaluation of ACS data, Bye developed methods for searching the Census Numident to find respondents' SSNs, successfully finding ACS respondents' SSNs about 93% of the time (Bye, 1999). With the SSN attached to the ACS record, it can then be linked to other data, particularly wage and salary data, for analysis[14].

A second function is to validate existing name/SSN/date of birth triples in the source administrative files. As described in previous sections, a validation strategy was developed, in

which various combinations of name transpositions, SSN transpositions, and date of birth errors are tested against the Census Numident in an attempt to validate these.

Results from a first analysis of this verification/validation and search algorithm are presented in table 4.

-- Insert table 4 about here --

As can be seen, about 875 million unduplicated person records, from all six StARS input files, were passed through the validation and search program. Of these, about 844 million were verified in the first phase, verification. About 29 million were passed to the search phase, and an additional 1.3 million SSNs were found. Thus, overall, 96.5% of the records passed into the program were validated. Instead of 1 in 10 SSNs being invalid, cited above in previous research, we have reduced that number in our database to about 1 in 20.

Privacy and Confidentiality Research, Protections, and Agreements

There are two levels of procedures to consider: The first is the committees established to take responsibility for forming policy and reviewing procedures; the second is the actual managerial and technical procedures established and implemented.

The Privacy Research Team (PRT)

An interdivisional group, called the Privacy Research Team (PRT), was established in 1994 to address Executive Staff's concerns privacy issues related to computer record matching and administrative records use in general, particularly given privacy implications for new technologies (Clark, 1997). From this beginning, the Privacy Research Coordinating Committee

[14] For the record, in their cover letter, respondents were informed that their records would be linked with other administrative data.

(PRCC) was established to coordinate privacy research and policy analysis as recommended by the Census Bureau's Policy Office (POL) in its 1998 Business Plan which was submitted to the Operating Committee, December, 1997, (Gates, 1998).

The Privacy Research Coordinating Committee

Thirteen members were selected by the associate directors from each directorate for the PRCC to represent their respective directorate's views. The adopted mission statement quotes "…[t]o assess the current state of privacy *practices* at the Census Bureau; to seek funding for a program of research, outreach, and education related to privacy; and to propose new integrated Bureau privacy policies," (Gates, 1998:1; italics theirs). To comply with the laws and regulations a review subcommittee was formed to monitor project proposals of the Administrative Records Research Staff's documentation of the project approval process and the ongoing inventory of active projects. This subcommittee is to make recommendations to the Project Review Board on an ongoing basis.

At their first meeting in April, 1998 they identified these priorities and activities to be addressed in the following twelve months:

- Public perceptions of Census's credibility regarding privacy and confidentiality.
- Policies regarding controversial privacy and confidentiality issues, studies, or surveys.
- Uniform understanding and agreement within Census on privacy and confidentiality policies.
- Technology's impact on privacy and confidentiality.

- Safeguards for protecting data privacy and confidentiality while maintaining the integrity of microdata files.
- Use of administrative records to augment or replace current data collection efforts.

Managerial and Technical Procedures - Restricted Access Policy for Systems of Administrative Records with Individual or Household Information

The Administrative Records Steering Committee, chaired by the Associate Director for Methodology and Standards, decided that data security and controlled access was of immediate and highest priority. Breaches of disclosure or unauthorized uses would destroy agreements with supplying agencies of data critical to Census programs and damage Census' reputation. Thus, the Restricted Access Policy for Systems of Administrative Records with Individual or Household Information was developed. The purpose of this policy is to protect the use and disclosure of administrative data that Census receives from other agencies (Clark, 1999). These protocols apply to internal access for sworn Census employees. The general rules or protocols are:

- Data are to be securely housed on special, stand-along computers at Census' Bowie Computer Center[15].
- Access is limited to a very small, select group of sworn employees including persons with Census Bureau Special Sworn Status.

[15] The Bowie Computer Center is a separate, secured, facility established in 1997 in Bowie, MD near the Maryland Science and Technology Center. It supports the large, Census-wide, computer applications.

- Access by persons who do not have sworn status is strictly prohibited which is part of all agreements with agencies supplying the data files.
- Data output from the Center is to be at the aggregate level only (subject to disclosure review) or stripped of all personal identifiers (e.g., name or SSN). Public notification will take place as required by the Privacy Act.

The Restricted Access Policy lays out specific managerial and technical controls to ensure the above rules or protocols.

Managerial Controls

These controls are directives to be carried out by personnel working with the administrative records under the management of the Administrative Records Research Staff, Project Research and Evaluation Division (PRED), Methodology and Standards Directorate of the U.S. Bureau of the Census.

Data Acquisition Policies

Acquisition of agencies' administrative records has been centralized within the Administrative Records Research Staff and coordinated by the Policy Office (Clark, 1999). When agreements are required by law or regulation, letters of agreement are often done at the Department Secretary level. However, in most cases agreements are memoranda of understanding (MOUs) between the two agencies. These MOUs undergo review by security and the legal offices, and they are signed at the Associate Director level or higher. The content delineates the authorized use, the legal authority, access restrictions, the retention period, and the disposition of the data at termination.

- The Internal Revenue Service (IRS) and the Social Security Administration (SSA) are the two main suppliers of administrative records used by Census.

Important acquisitions have been obtained after extensive negotiations; most important were the IRS's Individual Master Tax Return for 1997 and the Information Return Master File for 1996 which were acquired in October and April of 1998 respectively. The Social Security Administration's 100% NUMIDENT file was received by December, 1998. The NUMIDENT 2000 Dress Rehearsal Research file was received in February, 1998, (Leggieri, 1998).

These procedures were implemented in 1998.

Project Approval / Project Review

The existing Administrative Records Steering Committee, chaired by the Associate Director for Methodology and Standards, serves as an Administrative Records Review Board. This board reviews all projects for compliance to stated policy. Projects are to be compatible with Census mission/goals, policies, and data agreements. In addition, they must be of benefit to Census, methodologically feasible, and disclosure risk must be mitigated. Existing projects are "grand-fathered" for approval; projects that use only business administrative data are not subject to this review process.

These procedures were Implemented January, 1999.

System of Records Notification

This policy must be consistent with the 1974 Privacy Act, Computer Matching and Privacy Protection Act of 1988 and OMB Circular 1-130. System of Records public notification will be

made including changes in content or use of these systems. Responsibility for these notifications will be maintained by the Policy Office.

The Policy Office published the first System of Records in January, 1999

Linking of Administrative Records About People or Households

Administrative data that has personal identifying information (e.g., SSN, name) must be in a secured, restricted access, environment managed by the Administrative Records Research Staff (ARRS). Linkages with other administrative data or survey data will only occur in these restricted areas. Output files will be stripped of personal identifiers *unless* it has been approved by the Project Review Board and is in compliance with the supplying agency. For longitudinal studies, identifiers might be appended to maintain historical linkage.

Implementation was staged over eight months beginning in 1999 and files that were linked before November 15, 1998 were migrated to the ARRS system in the eight month transition period.

Security Officer

The Assistant Division Chief for Administrative Records will designate a security officer who will:

1) serve as the liaison with the Census Security Office to develop and maintain security plans and verification of security protection and access restrictions;
2) coordinate and monitor user training for file security procedures; and,
3) monitor the office and processing activities for compliance.

Implementation took place prior to 1999.

Employee Awareness and User Notification

Census will officially inform and caution employees who have access to administrative records, including persons who have Census Bureau Special Sworn Status, of the legal, regulatory, and other requirements to insure confidentiality is maintained including penalties for unlawful disclosure.

Division chiefs and all division users must be notified of disclosure limitations and associated penalties upon release of any output files from the restricted environment. The requesting manager must sign an acknowledgement of the receipt of the files and of disclosure limitations. The following are user notification procedures:

- As required by Title 13, a sworn affidavit of nondisclosure that applies to confidential information from Census's collections and administrative records will be signed when duty begins.
- The Director will send an annual reminder to all employees about their obligation under the affidavit.
- The Director will send an annual reminder to all employees about the confidentiality provisions that apply to tax data supplied under authority of the Internal Revenue Code and its regulations including penalty provisions.

- The Economic Programs Directorate will send an annual notice to its employees about the confidentiality and penalty provisions of the IRS "official use only" information which describes the content and format of the supplied tax records.

User notification was completed for all files was completed in June of 1999.

Disclosure Review

Any publication of statistical products are subject to disclosure avoidance procedures proscribed by the Disclosure Review Board. If required by the supplying agency, any product based in part or in whole on administrative records data may be subject to additional disclosure review.

Implementation took place prior to 1999.

Technical Controls

Technical controls address the physical and technological features of confidentiality protection through limited access. These controls are intended to focus on how the administrative data are actually used in the technical environment and to govern the physical and information technology security.

Computers

The Administrative Records Research Staff (ARRS) hardware system is the at the center of the restricted access protocols which is under the direction of the Census Bureau Security Office. The hardware system, designed to handle a very large amount of data, consists of 2 DEC Alpha 8400 processors with 3 Terabytes online disk storage and 18 Terabytes nearline disk storage. It is housed at the Bowie Computer Site. Access must comply with the Computer Security Act of 1997 and the DOD 12/85 report referred to on page 9.

Implementation is complete and the security plans for systems were in place on November 1, 1998.

Users / Passwords

Access is limited to persons with the "need to know." Levels of access include (0 = highest level of access):

Level 0 System Administrator, Security Officer, and primary person(s) in charge of the file and the people with access to entire microdata file are permitted access to the microdata on the secured, restricted access, ARRS computer system.

Level 1 Persons with access to extracts of microdata are permitted access to the microdata on the secured, restricted access, ARRS computer system.

Level 2 Persons with access to non-disclosure-proof summary data and microdata extracts with no personal identifiers, but these persons are not allowed access to the restricted ARRS computer system.

Level 3 Persons with access to disclosure-proof summary data only, but these persons are not allowed access to the restricted ARRS computer system.

Following the rules laid out in the security plan, only Levels 0 and 1 personnel, who are usually the Administrative Records Research Staff (ARRS), are permitted to access the secured, restricted access, ARRS computer. However, the Assistant Division Chief for Administrative Records Research can approve level 1 access to others on a case by case basis. Level 2 personnel

may use administrative data on computers that meet the Census Bureau's and supplying agencies' requirements.

In accordance with Administrative Memorandum 9, "Guidelines for Census Confidential Material", all access screens and print and other media (e.g., tapes, etc.) contain statements that disclosure is prohibited and that data are to be used for authorized purposes only.

Implementation of these controls was completed by November 1, 1998.

Telecommunications

Communication with the restricted access system will use either point-to-point dedicated lines or encryption. The telecommunications office is to encrypt all lines between the Bowie Computer Center, Suitland Headquarters Offices, Washington Plaza, and the outlying (regional, etc.) offices.

Implementation was to be completed by December 31, 1998 except for the outlying offices which was to be completed by April 30, 1999 (if funds were available).

Audit Trail

Logs of user access are generated automatically and routinely reviewed by the ARRS Security Officer. These rules meet the Census IT Security Handbook guidelines.

Implementation started November 1, 1998.

Printed Materials

Any materials containing personal identifiers or other nonpublic information for which disclosure is prohibited will be kept in locked areas and are to be destroyed in accordance with the Census Administrative Memorandum General 16: "Destruction of Sensitive Materials."

These procedures were in place prior to December 1, 1998.

Electronic Storage

All CD ROMs or computer tapes containing restricted data will be marked as sensitive information and kept in locked areas. The data are destroyed by degaussing and reformatting data on tapes and by breaking and destroying CD ROMS in accordance with the "Destruction of Sensitive Materials" guidelines.

Removal of Social Security Numbers

Most importantly, the ARR staff translates incoming SSNs into Personal Identification Keys (PIK) in place of SSNs which is an encryption scheme that is very difficult to break without the originating table. Except for ARR staff, anyone working on these data would only see PIKs, not the SSNs.

Administrative Records Challenges

Where do administrative records data come from?

Figure 5 illustrates the process of data collection in the administrative records context (taken variously from Ma, 1986; Redman, 1992; and Judson and Popoff, 1998):

-- Insert Figure 5 about here --

In this figure we have identified how records are created and used, and, at each step, where there is potential for error. A record begins as an event or an object in the Areal world@ outside the database. Some of these are actually identified (note already that some events and objects that are *really there* do not get identified) as observed events and objects. Some of these observed events and objects get recorded (again, some do not) and become part of the administrative record. These are then arranged in some form of database. Later, when the analyst approaches these records, he or she develops analyses (Aqueries@ in the language of data mining), and makes presentation of the results. Finally, using either the usual statistical framework or policy-making framework, we presume that some decision is made.

At each step in this process there is potential for error or misinterpretation. At the event/object level, policy changes can change the definition of an event or object (as when AFDC ceased to exist; or when a mobile home is recoded from Areal@ property to Apersonal@ property). As events and objects become observed, the Aontologies@ (described in more detail below) which observers use to categorize the world enter in; likewise, some events and objects do not reach a threshold for official notice (as in the deviance literature where the Aofficer on the beat@ makes the call whether to identify an event as a problem or not; see, e.g., Laudon, 1986; Light, 1990). Some of those observed events become recorded, and there is potential for data entry errors and limitations of coding schemes for mapping the Areal world@ events into their database representations. As data are moved from place to place in computer systems, data management problems (corruption in transit, changes in formatting, etc., documented in Stevens, Richmond, Haenn, and Michie, 1992:171-178) potentially arise. Finally, as the analyst extracts data using queries of various kinds, he or she may discover that the query is either syntatically incorrect in a nonobvious way, or the query generates results whose structure itself is spurious[16].

---

[16] From a recent discussion of AFDC/TANF: "One may be interested in examining demographic, employment and other characteristics of persons using AFDC/TANF. Because AFDC/TANF is a means-tested cash benefit, it is treated as a source of income for the recipient (as well as counting towards total family and household income). It is certainly true, in the case of AFDC/TANF, that the benefit is intended for more persons than just the recipient. Nevertheless, one may be interested in the characteristics of those applying for AFDC/TANF and/or the characteristics of those persons covered under the program. Characteristics, unsurprisingly, can change dramatically depending upon which frame you choose for your analysis. For example, if one were to examine the gender of those persons covered by AFDC/TANF, about 45 percent would be male. However, if one's frame is recipients (i.e. persons who apply and receive AFDC/TANF as income), then about 5 percent are male."

Similarly, in a comparison of participation rates in benefits programs: "When judging usage of any program, there is a distinct difference between what are often called "participation" and "take-up" rates. Both measures are essentially ratios (or percentages) of those using a particular program to some "relevant" population. Participation rates use the entire population as the base while take-up rates use the subset of the population that meet eligibility requirements for the program. This subtle difference in the denominator can yield vastly different interpretations of policy efficacy. For example, imagine that we see a trend of lower

Ontologies: An Approach to Data Quality

The term "data quality" is in itself a challenging entity. For example, what does it mean for a measured data element to be "right?" How "wrong" is "too wrong?" Let us illustrate with a simple example, involving the simplest of elements: A person's sex. At first glance, the quality of a particular measurement of that variable seems obvious. Either we have ascertained a person's sex correctly or we have done so incorrectly. However, if we take this simple view, we forget that, as with all categories, the categories of "male" and "female" are in part socially constructed. For example, what about a hermaphrodite? Or an individual with Klinefelter's syndrome (an XXY sex chromosome pattern rather than the standard XX or XY pattern)? If the database were a medical database focusing on genetic anomalies, the simple categories of "male" and "female" are no longer so easy to ascertain, and it might be of great importance to know this precisely. Similarly, if our database is for a state's correctional system, ascertaining the "sex" of a client is not necessarily a trivial matter[17].

---

food stamp caseloads. Whether we interpret this as "good" news may depend upon the relative changes in participation and take-up rates. If, in these times of falling caseloads, participation rates are lower while take-up rates remain unchanged, this is evidence that the fall in caseloads is attributed to the fact that fewer households are meeting eligibility requirements. One might conclude that having fewer eligible households is a sign of generally better economic times. Conversely, if the drop in food stamp caseloads coincides with stable participation rates, this suggests that take-up rates must have fallen. One might argue that these decisions not to participate, while still remaining eligible, are not a strong indicator of economic improvement."

[17] Judson and Sigmund (1995), note that the Oregon Department of Corrections database contains codes for "hermaphrodite" and "unknown" sexes.

We can easily go further: If our database is focused on social science topics, it is perhaps better to focus on the individual respondent's gender, rather than their sex. However, we know that "gender" itself is socially constructed and has variations and nuances that are not easily, nor appropriately, captured in a form with two boxes, "Male" and "Female."

Finally, let us consider the *mutability* of the data element. Even if we wish to take a fairly strict definition of sex, *in a population database* we are forced to consider individuals whose sex changes over a period of time.

Every comment above can be multiplied about every conceivable data element--and we must firmly keep in mind that social definitions and the uses of a database are changing, and that when a coder (either agency personnel or an individual respondent) does not have the categories that are appropriate to describe him or herself, they will most likely choose the best fitting. Choosing the "best fitting" response, rather than choosing the "right" one, is itself constructing social reality, not necessarily reflecting it.

Because even the simplest data elements run into definitional questions, it follows that the notion of "data quality" must follow from what database experts refer to as an ontology (Hovy, 1997; Wand and Wang, 1996). At its basic level, the branch of philosophy known as ontology is the study of what is--the nature of reality. For a computer database expert, an ontology is a method for encoding the "real world" in a computer representation. Wand and Wang provide four diagrams that represent this

encoding and identify failures (such that we can say that data quality is a problem). Figure 6 describes these encodings.

-- Insert figure 6 about here --

In this figure, the square boxes represent the states of nature, as it were: What the data would say if we were able to completely characterize the ontology and an object=s attributes in that ontology. The rounded boxes represent the computer representation we choose.

In the framework proposed by Wand and Wang, the key capability of a database system is the ability to *reconstruct the original states of nature*. That is, if we begin with the rounded boxes on the right, we should be able to completely and unambiguously establish which square box on the left the object came from. In the upper left hand corner of the diagram is a Aproper representation@--even though an object is encoded from state 3 (in the real world) to state 3 or 4 in the database, we still can unambiguously recreate the real world state from the database information. In other words, whether our database says the object is in state 3 or in state 4, we know that it belongs to state 3 in the real world. The remaining corners represent failures of different kinds. In the lower left hand corner we have ambiguous representation--in a nutshell, if we know that the object is in state 2 in our database, we cannot reconstruct whether it started in state 3 or in state 2 in the real world. We can=t Amap back@ to the real world from our database. In the upper right hand corner, we encounter incomplete representation--we have no category for an object in state 3 in the real world. There is no way to encode, so if we are forced to do so by our data collection responsibilities, we must miscode the object in state 3. In the lower right

hand corner, we encounter meaningless states--we have codes in the database with no real world counterparts. Thus, if we find a Astate 3@ object in our database, we have no way of knowing where it came from and no way to determine if there is anything in the real world that needs to be reconstructed. Thus, whenever one develops a model that maps from one categorization scheme to another, this model can be considered the mathematical equivalent of a "translation." And as is well known in language translation, not all translations are perfect, even with a perfect translator. Some concepts that can be well-expressed in one "language" can be only imperfectly expressed in the other.

Ontological Challenges

The problem database ontologies create in our context are clear: Our goal is to construct a roster of census persons at census addresses, and the ontologies that the Census Bureau uses do not always match those of data suppliers.

For example, a delivery address suitable for receiving a payment check may not suffice for putting individuals at a street address for purposes of geographic assignment. (The most obvious examples are post office boxes and commercial mailing services.) It is often difficult to distinguish individual units within the Basic Street Address[18]. That is, it is sometimes hard to determine if an address refers to different apartment units at one Basic Street Address. As a result, making standard demographic measures (such as the calculation of persons per household) is not always a straightforward process.

Additionally, we note that race coding itself is a database ontology problem. As is well known, according to Census Bureau and OMB definitions, "Hispanic origin" is an ethnicity and not a race. However, "Hispanic origin" is treated as a separate race on the SSA Numident--thus, *from the point of view of the SSA Numident*, Hispanic is a *race group.* Prior to 1980, this race group did not exist as a selection option on the SSA form; *hence a person born before 1980 could not be Hispanic*[19]. The race model that we have described in previous sections effectively "translates" from one ontology to the other. The fact that it is a probabilistic translation (that is, the model is a probabilistic model and not a deterministic one) is only a side detail.

A further conceptual distinction between databases in our context is that *transaction data are not equivalent to person data*. For example, both the SSA Numident and the HIS files are transaction files, and a record represents an interaction between the source agency and our target, the target person. Obviously, some characteristics of a person can change between one transaction and the next, even if no data entry error has occurred. For some concepts, such as residence or mailing address, this is obvious. For others, however, it is less obvious—one's race can change over time as one self-identifies differently given different motivations. Or, one's sex can change. Or the coding scheme recording the data can change, and a person effectively moves from one class to another, such as the SSA race categories demonstrate.

[18] In our context, a "Basic Street Address" is defined as a street number/street name combination that can be located on a map.

Related to the change of characteristics is the fact, noted in Blumberg and Goerman, 2000, that naming conventions vary across cultures. How many names does a person have (and in what order)? It depends on their culture, marital status, lineage status, even their personal preference. And to whom does the obtained administrative data record apply? While the target person for the agency might be a particular client, our legal system allows for proxies to interact with the administrative agency in the place of the client. This is particularly true for Medicare and IRS records. Consider the following address:

John Wilson

C/O Mary Wilson

1004 Laurel Lane

Rockmount, MD 22345

In this example, it would seem clear that the address applies to Mary Wilson. John Wilson may or may not live at this address.

Addresses that are difficult to place on the ground

Huang and Kim (2000), in their 1% sample study, noted that about 10% of the addresses in their file were rural style. Rural style addresses are exceptionally difficult to place in a particular geographical location without the benefit of map spotting and physical location information. Further, they found that Post Office Boxes accounted for 45% of IHS, 9.5% of Medicare, 7.5% of IRS 1040, 6.8% of SSS, 3.8% of IRS 1099, and .4% of HUD-

[19] It is important to be emphatic on this point: *From the point of view of the database*, the racial/ethnic

TRACS addresses. Sater (1995), in a match of IRS to Current Population Survey addresses, noted that 86.5% of tax return cases had the same address as the CPS residence address; further, 94% coded to same county. So, while we are mostly able to identify a county of residence, this does not mean that we can place persons in particular subcounty places. An example will suffice:

John Smith
H&R BLOCK
P.O. Box 12
Greenway, MD 29752

A naïve treatment of this record would indicate that John Smith lives at H&R Block. Obviously this cannot be the case. However, are we sufficiently confident that Mr. Smith lives in Greenway? In the county in which 29752 falls? In Maryland? In an enumeration or an estimation context, where shall we place Mr. Smith?[20]

Addresses with both business and residential components

A further complication of our address handling is the fact that many addresses have both business and residential components. Thus, even if we are able to identify an address as a commercial address (in fact, methods are in place to do so), we cannot be certain that that address does not also house persons. For example:

Dean H. Judson

---

group *did not exist*.

[20] Currently, we refer to such persons as "floaters" because we cannot place them on the ground. Methods for handling "floaters" have not yet been devised. Obviously, they might very well be treated differently in an enumeration context versus an estimation context—while we might be willing to place them on the ground using some probabilistic or distributive estimation system, it seems doubtful that such an idea would be acceptable in an enumeration context.

JUDSON OLD GROWTH LOGGING, STRIP MINING & SPOTTED OWL

EXTERMINATION SERVICES

45850 Backwoods Highway

Boondocks, OR 97701

This address is obviously commercial. Yet, does it house a camp? Do migrant and seasonal workers reside there? Does Dean H. Judson reside there or just receive mail there?

Unduplication and matching

When addresses or personal characteristics are measured with substantial variation, it is often not obvious whether a particular pair of records represent a duplicate or not. Yet, with multiple files being combined, unduplication decisions *must* be made.

-- Insert tables 5 and 6 about here –

As can be seen in the tables 5 and 6, when attempting to unduplicate addresses, many different inferences can be made. When attempting to update a list such as the Master Address File or the Standard Statistical Establishment List, how can we determine whether a new record represents a duplicate or not?

Variations in data from different sources

Huang and Kim (2000), in their 1% sample study, found that of the 50% of SSNs found on multiple files, about 1% have more than one gender recorded; about 32% have multiple addresses; and about 2% have multiple races. As averred above, these variations

must be handled—we are not free to abstain from making a decision. For example, consider the decision that must be made regarding the following address comparison:

Sam Smith	Sam Smith
Box 2 Rural Route 37	486 Main St
Westport, VA 32784	Fairfield, VA 33412
(Dated 10/14/98 from Medicare)	(From TY97 IRS file, filed sometime in 1998)

Suppose today is January 1, 1999. The address in the first column is much harder to place on the ground, because it is rural style. However, the timing of the address is the second column is much more uncertain. Where does Mr. Smith live on 1/1/1999? Again, in both the enumeration and the estimation contexts, *we are not free to abstain from this decision*.

Limited and inconsistent microdata content

As noted in earlier sections, many files have limited microdata content. For those that are found on the Numident, we can "impute" or "model" microdata from the approximately equivalent Numident fields. *All* files in the StARS system suffer from this limitation, and *all* fields have some percentage of missing and inconsistent data.

Despite the limitations of the source data files, we are able to impute or model some microdata content. However, it is an open question whether imputed or modeled race, performed on a wide scale, will be acceptable in the enumeration context. (It seems likely that modeled race would be acceptable in an estimation context.)

Changing information states

The problem of ongoing administrative records databases is a distinct problem from "point in time" data collection, in that the target population *itself* is changing over time. The databases are attempting to track this moving target, some with greater and some with lesser success. A datum might be *correct* at one time and *incorrect* at another, or, even worse, *a datum might be correct for one purpose and incorrect for another purpose*. Because of this "tracking" feature, information states change over time and over databases: For example, for a particular record representing a person of advanced age, we must ask: Is the person alive or not? The answer depends upon whether some administrative database has had contact with them recently. Obviously, if the record keeper heard from them yesterday, there is a higher probability that they are still alive than if the record keeper heard from them forty years ago. Similarly, for a record with different addresses on two databases, which address is the more correct--the most recent address or the most complete address? Or, further, which is more correct, the physical address or the mailing address? This last example illustrates how a datum might be correct for one purpose and incorrect for another: A post office box or rural route might be perfectly acceptable for delivering mail, yet completely unacceptable for allocating persons to blocks for creating redistricting data.

It is clear, from these examples and the related ongoing research, that one database provides information about the other, provided that matching can be performed; however,

it is very important to remember that the data processing requires complex, and *substantively important*, decision logic at each step.

Conclusion

We must continue to be vigilant against the naive view that administrative records data are the "truth." Further, we must be exceptionally careful when a database administrator claims that his/her database "has" some datum. We must ask questions such as:

- With what coding scheme is the datum collected?
- To what extent does that datum cover your client population?
- To what extent does that datum cover *our* target population, which might not be the same as yours?
- With what degree of accuracy is the datum collected?
- Over what time frame is the datum collected?
- Does the datum change over time?
- When is the datum processed so that I can use it?
- Is it legal for me to obtain this particular datum and how must I protect it?

Such questions are not commonly discussed in the typical "point in time" data collection texts, yet they are fundamental to administrative records research.

Nonetheless, despite these challenges, administrative records research forges ahead at the U.S. Census Bureau. Much has already been learned about coverage, data errors, translation and data handling, and potential uses of these databases. When the current

experiments come to their conclusion, we will know much more about what questions to ask and what answers to expect.

References

Abraido-Lanza, Ana F., Dohrenwend, Bruce P., Ng-Mak, Daisy S., and Turner, J. Blake (1999). The Latino mortality paradox: A test of the "salmon bias" and healthy migrant hypotheses. American Journal of Public Health, 89:1543-1548.

Agresti, Alan (1996). An Introduction to Categorical Data Analysis. New York, NY: John Wiley and Sons.

American Statistical Association (1977). Report of the Ad Hoc Committee on Privacy and Confidentiality. The American Statistician , 31, 59-78.

Anderson, Margo J., and Fienberg, Stephen E. (1999). Who Counts? The Politics of Census-Taking in Contemporary America. New York, NY: Russell Sage Foundation.

Bell, William (1992). Using Information from Demographic Analysis in Post-Enumeration Survey Estimation. Research Report 92/04, U.S. Bureau of the Census.

Biemer, Paul (2000). Triple System Estimation with Erroneous Enumerations. U.S. Census Bureau internal memorandum dated March 6, 2000.

Bishop, Yvonne, Feinberg, Stephen, and Holland, Paul (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press.

Blumberg, Rae, and Goerman, Patricia (2000). Family Complexity Among Latino Immigrants in Virginia: An Ethnographic Study of their Households Aimed at Improving Census Categories. Draft final report for the "Complex Households and Relationships in the Decennial Census and Ethnographic Studies Project." Washington, D.C.: U.S. Census Bureau.

Brackstone, George J. (1988). Statistical Uses of Administrative Data: Issues and Challenges" in Coombs, J. W. and Singh, M.P. (Eds.), Statistical Uses of Administrative Data : An International Symposium. Ottawa: Statistics Canada.

Brown, L., Eaton, M., Freedman, D., Klein, S., Olshen, R., Wachter, K., Wells, M., and Ylvisaker, D. (1999). Statistical controversies in the Census 2000. To appear in Jurimetrics.

Bye, Barry V. (1997). Administrative Records Census for 2010 Design Proposal. Rockville, MD: Westat, Inc.

Bye, Barry V. (1998). Race and Ethnicity Modeling with SSA Numident Data. Administrative Records Research Memorandum Series #19, U.S. Census Bureau.

Bye, Barry V. (1999). Social Security Number Search And Verification At The Bureau Of The Census: American Community Survey and Other Applications. Administrative Records Research Memorandum Series #31, U.S. Census Bureau.

Bye, Barry V., and Thompson, Herbert (1999). Race & Ethnicity Modeling w/SSA Numident Data: Two Level Regression Model. Administrative Records Research Memorandum Series #22, U.S. Census Bureau.

Chao, Amy, and Tsay, Paul (1998). A sample coverage approach to multiple-system estimation with application to census undercount. Journal of the American Statistical Association, 93: 283-293.

Childers, Danny (2000). Accuracy and Coverage Evaluation: The Design Document. DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1.

Childers, Danny, and Hogan, Howard (1984). Matching IRS records to Census records: Some problems and results. Proceedings of the Section on Government Statistics. Alexandria, VA: American Statistical Association.

Clark, Cynthia Z.F. (1997). Confidentiality and Privacy Research Status. [Memorandum]. Suitland, MD: US Bureau of the Census.

Clark, Cynthia Z. F (1999). Restricted Access Policy for Administrative Records. Internal Census Bureau Memorandum dated June 25, 1999. Suitland, MD: US Bureau of the Census.

Coder, John (1992). Using administrative Recoding Information to Evaluate the Quality of the Income Data Collected in the Survey of Income and Program Participation. Proceedings of Statistics Canada Symposium 92: Design and Analysis of Longitudinal Surveys. Ottawa: Statistics Canada.

Coder, John, and Scoon-Rogers, Linda (1995). Evaluating the Quality of Income Data Collected in the Annual Supplement to the March Current Population Survey and the Survey of Income and Program Participation. SIPP Working Paper Series, No. 215, U.S. Bureau of the Census, Washington, D.C.

Copas, J.B., and Hilton, F.J. (1990). Record linkage: Statistical models for matching computer records. Journal of the Royal Statistical Society, Series A, 153:287-320.

Czajka, John (1999). Can we count on administrative records in future U.S. Censuses? Presentation at the Bureau of the Census, December 15, 1999.

Czajka, John L., Moreno, Lorenzo, and Schirm, Allen L. (1997). On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population. Washington, D.C.: Mathematica Policy Research.

Darroch, John, Feinberg, Stephen, Glonek, G., and Junker, Brian. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. Journal of the American Statistical Association, 88:1137-1148.

David, Martin, Little, Roderick J.A., Samuhel, Michael E., and Triest, Robert K. (1986). Alternative methods for CPS income imputation. Journal of the American Stastistical Association, 81:29-41.

Dempster, A., Laird, N., and Rubin, Donald B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39:1-38.

Department of Health and Human Services, Office of Inspector General (1990). Extent of Social Security Number Discrepancies. Washington, DC: Department of Health and Human Services.

Edmonston, Barry and Schultze, Charles (Eds.) (1995). Modernizing the U.S. Census. Washington, D.C.: National Research Council.

Ericksen, E., Kadane, John, and Tukey, John (1989). Adjusting the 1980 Census of Population and Housing. Journal of the American Statistical Association, 84: 927-944.

Falkenstein, Matthew, Resnick, Dean R., and Judson, Dean. H. (2000). The Mortality Module of the Statistical Administrative Records System. To be completed in the Administrative Records Memorandum Series, U.S. Bureau of the Census.

Fellegi, Ivan P., and Sunter, Alan B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64: 1183-1210.

Gates, Gerald W. (1998). Privacy research coordinating committee: Yearly progress report, April, 1998 – April, 1999. Internal report. Suitland, MD: U.S. Bureau of the Census.

Government Organization and Employees, The Freedom of Information Act, 5 U.S.C. § 552, as amended, 1997.

Government Organization and Employees, The Privacy Act of 1974, 5 U.S.C. . § 552a as amended, 1997.

Hogan, Howard (2000). The Accuracy and Coverage Evaluation: Theory and Application. Paper presented at the 2000 Joint Statistical Meetings, Indianapolis, Indiana, August, 2000.

Hogan, Howard, and Robinson, J. Gregg (1993). What the Census Bureau's Coverage Evaluation Programs Tell us About Differential Undercount. Paper presented at the 1993 Research Conference on Undercounted Ethnic Populations, Richmond, VA, May 5-7, 1993.

Hovy, Eduard (1997). A standard for large ontologies. Information Sciences Institute [Online]. Available: http://www.isi.edu/nsf/papers/hovy2.htm [1997, December 29].

Huang, Elizabeth, and Kim, Jay (2000). One Percent Sample Study Report. Administrative Records Research Memorandum Series #42, U.S. Census Bureau.

Inmon, William H. (1996). Building the Data Warehouse, Second Edition. New York, NY: John Wiley and Sons, Inc.

The Internal Revenue Service 26 U.S.C. § 7213.

The Internal Revenue Service 26 U.S.C. § 7431.

Judson, D. H. (1992). Performing Loglinear Analysis of Cross-Classifications. Stata Technical Bulletin, 6:7-17.

Judson, Dean H. (2000a). Use of Administrative Records for Triple System Estimation. U.S. Census Bureau Internal Memorandum, March 15, 2000. Washington D.C.: U.S. Census Bureau.

Judson, D.H. (2000b. A partial order approach to record linkage. Paper to be presented at Los Alamos National Laboratories seminar series.

Judson, D.H. (2000c). Estimating Fellegi-Sunter match weights using logistic regression. Unpublished paper.

Judson, D.H., and Popoff, Carole L. (1996). Assessing the quality of data in data warehouses and administrative records: A research and statistical approach to data quality [Online]. Available at the Workshop on Research and Development Opportunities in Federal Information Services, http://www.isi.edu/nsf/agenda.html [1998, January 3].

Judson, D.H., Popoff, Carole L., and Batutis, Michael (2000). An administrative records approch to evaluating of the accuracy of U.S. Census Bureau County Population Estimates. Proceedings of the 6th Annual Applied and Business Demography Conference. Bowling Green: Bowling Green State University.

Judson, D.H., Popoff, Carole L., and Batutis, Michael (2000). An evaluation of the accuracy of U.S. Census Bureau County Population Estimates. Paper submitted to Statistics in Transition, under review.

Judson, Dean H., and Sigmund, Charles L. (1995). The Shared Information System, Year Two: An Evaluation of the Uses of the Mobility Continuum Concept for the Workforce Quality Council. Reno, NV: Decision Analytics.

Killion, Ruth Ann (1999). 1999 Administrative Records File Acquisition. Administrative Records Research Memorandum Series #37, U.S. Census Bureau.

Kim, Myoung Ouk, and Sater, Douglas (2000). Defining the Medicare Data Universe for the U.S. Census Bureau's Population Estimates Program. U.S. Bureau of the Census Internal Memorandum, August 29, 2000. Washington D.C.: U.S. Bureau of the Census.

Knott, Joseph J. (1994). Proposed Uses of Administrative Records in the 1995 Census Test. U.S. Census Bureau Internal Memorandum, March 14, 1994. Washington D.C.: U.S. Census Bureau.

Lahiri, Partha, and Larsen, Michael D. (2000). Model-Based Analysis of Records Linked Using Mixture Models. Paper presented at the 2000 Joint Statistical Meetings. Indianapolis, Indiana, August 8-13, 2000.

Larsen, Michael (1999). Predicting the Residency Status for Nonmatching Administrative Records. Administrative Records Research Memorandum Series #28, U.S. Census Bureau.

Laudon, Kenneth C. (1986). Data quality and due process in large interorganizational record systems. Communications of the ACM, 29:4-11.

Leggieri, Charlene A. (1998). 1997 Administrative Records File Acquisition. Administrative Records Research Memorandum Series #18. Suitland, MD: U.S. Bureau of the Census.

Leggieri, Charlene, and Killion, Ruth Ann (2000). Administrative Records Experiment in U.S. Census 2000. Unpublished document available from the U.S. Census Bureau.

Light, Stephen C. (1990). Measurement error in official statistics: Prison rule infraction data. Federal Probation, 54:63-68.

Long, John F., and Wetrogan, Signe I. (1990). *Creating Annual State-to-State Migration Flows with Demographic Detail by Merging Census Survey and Administrative Records Tabulations*. Current Population Reports, Series P-23, No. 166. Washington D.C.: U.S. Bureau of the Census.

Ma, Juliana Mei-Mei (1986). A Modeling Approach to System Evaluation in Research Data Management. Ph.D. Thesis, University of North Carolina, Chapel Hill, NC.

McLachlan, Geoff, and Krishnan, T. (1997). The E-M Algorithm and Extensions. New York: John Wiley and Sons, Inc.

Marquis, Kent, Wetrogan, Signe, and Palacios, Henry (1996). Towards a U.S. Population Database from Administrative Records [Online]. Available at the Working Papers in Survey Methodology, http://www.census.gov/srd/papers/pdf/km9601.pdf, [2000, September 18].

Miller, Esther, Judson, Dean H., and Sater, Douglas (2000). The 100% Census Numident: Demographic Analysis of Modeled Race and Hispanic Origin Estimates Based Exclusively on Administrative Records Data. Presented at the 2000 meetings of the Southern Demographic Association, New Orleans, LA.

Moore, Jeffrey C., Stinson, Linda L., and Welniak, Edward J., Jr. (2000). Income measurement error in surveys: A review. In Sirken, M., D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (Eds.), Cognition and Survey Research. New York, NY: John Wiley and Sons.

Mulry, Mary, Davis, Mary, and Hill, Joan E.. (1997). A Study of Heterogeneity of Census Coverage Error for Small Areas. Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Myrskyla, Pekka (1991). Census by questionnaire--Census by registers and administrative records: The experience of Finland. Journal of Official Statistics, 7:457-474.

Myrskyla, Pekka, Taeuber, Cynthia, and Knott, Joseph (1996). Uses of administrative records for statistical purposes: Finland and the United States. Unpublished document available from the U.S. Census Bureau.

Nelson, Charles (1985). Adjusting Imputed Interest Amounts Based on Results of the CPS-IRS Exact Match. memorandum for John Coder, Chief, Income Statistics Branch, Population Division, Bureau of the Census, February 7, 1985.

Neugebauer, Sharon, Perkins, R. Colby, and Whitford, David C. (1996). 1995 Census Test Results: First stage evaluations of the 1995 Census Test Administrative Records Database. 1995 Census Test Results Memorandum 41, March 14, 1996, U.S. Census Bureau.

Pearson, Lucinda S. and Douglas K. Sater (1999). A New Approach to Migration Data Production from Administrative Records: Return-Based vs. People-Based. Paper presented at the 1999 Southern Demographic Association meetings, San Antonio, TX, Oct. 28-30.

Pistiner, Arona (1999). Administrative Records Census Experiment In 2000 (AREX 2000). Internal Census Bureau memorandum.

Prevost, Ron (1996). Administrative Records and the New Statistical Era. Paper presented at the 1996 Annual meeting of the Population Association of America. New Orleans, LA, May 9-11, 1996.

Prevost, Ron (1997). The Usefulness of IRS Information Returns in the Development of a National Administrative Records Database. Administrative Records Research Memorandum Series #12, U.S. Census Bureau.

Prevost, Ron, and Leggieri, Charlene (1999). Expansion of Administrative Records Uses at the Census Bureau: A Long-Range Research Plan. Paper presented at the November 1999 Meeting of the Federal Committee on Statistical Methodology, Washington D.C.

Rasch, George (1980). Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: Univ. of Chicago Press.

Redfern, Phillip (1989). European experience of Using Administrative Data for Censuses of Population: The Policy Issues that Must be Addressed. Survey Methodology, 15:83-99.

Redman, Thomas (1996). Data Quality for the Information Age. Norwood, MA: Artech House.

Rubin, Donald B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

Sailer, Peter, Weber, Michael, Yau, E. (1993). How Well Can IRS Count the Population? Proceedings, Government Statistics Section, American Statistical Association. Alexandria, VA: American Statistical Association.

Scheuren, Fritz, and Winkler, William E. (1993). Regression analysis of data files that are computer matched. Survey Methodology, 19:39-58.

Scheuren, Fritz, and Winkler, William E. (1997). Regression analysis of data files that are computer matched - Part II. Survey Methodology, 23:157-165.

Short, Kathleen, Garner, Thesia, Johnson, David, and Doyle, Patricia (1999). Experimental poverty measures: 1990 to 1997. U.S. Census Bureau, Current Population Reports, P60-205. Washington, D.C.: U.S. Government Printing Office.

Stevens, David W., Richmond, Peggy A., Haenn, Joseph F., and Michie, Joan S. (1992). Measuring Employment Outcomes Using Unemployment Insurance Wage Records. Washington, D.C.: Research and Evaluation Associates, Inc.

Sweet, Elizabeth (1997). Using Administrative Record Persons in the 1996 Community Census. Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Taueber, Cynthia, Lane, Julia, and Stevens, David (2000). The Why, What, and How of Converting Program Records and Summarized Survey Data to State and Community Information Systems. Paper presented at the Conference, Developing Public Policy Applications with Summarized Survey Data and Community Administrative Records. Baltimore, MD, June 6-7, 2000.

Thompson, Herbert (1999). The Development of a Gender Model with SSA Numident Data. Administrative Records Research Memorandum Series #32, U.S. Census Bureau.

U.S. Census Bureau (1998). Restricted Access Policy for Systems of Administrative Records with Individual or Household Information. Internal Census Bureau Memorandum dated December 31, 1998. Suitland, MD: U.S. Census Bureau.

U.S. Census Bureau (1999). 1999 StARS Development and ARR File Management Flowcharts, U.S. Census Bureau internal document.

U.S. Census Bureau (1999). 1999 Census Numident Programming Specification. Administrative Records Research Staff: Specifications & Operations Area, catalog number CNUM9901-01, September 23, 1999.

U.S. Census Bureau (2000). Social Security Number Verification Programming Specification: Social Security Number Verification Against the Census Numident, Second Draft. Administrative Records Research Staff: Specifications & Operations Area, catalog number: STAR9903-01, August 4, 2000.

U.S. Census Bureau (2000). StARS Person Edit Programming Specifications: StARS 1999 Development Person Edit General Specifications, Administrative Records Research Staff: Specifications and Operations Area, catalog number STAR9904-00, April 27, 2000.

Wand, Yair, and Wang, Richard Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39: 86-95.

Weidman, Lynn, and Alexander, Charles (1999). Estimation for the American Community Survey: Ongoing work, planned work, and issues. Paper presented to the Census Advisory Committee of Professional Associations Meeting, October 21-22, 1999.

Winkler, William (1995). Matching and record linkage. In: B.G. Cox, et. al., Eds., Business Survey Methods. New York, NY: John Wiley.

Winkler, William (2000). Machine Learning, Information Retrieval, and Record Linkage. Paper presented at the 2000 Joint Statistical Meetings, Indianapolis, IN, August 15-18, 2000.

Wolter, K. (1986). Some coverage error models for census data. Journal of the American Statistical Association, 81: 338-346.

Zanutto, E. (1996). Estimating a population roster from an incomplete census using mailback questionnaires, administrative records, and sampled nonresponse followup. Presentation to the U.S. Bureau of the Census, 8/26/96.

Zanutto, Elaine, and Zaslavsky, Alan M. (1996). Estimating a population roster from an incomplete census using mailback questionnaires, administrative records, and sampled nonresponse followup. In Proceedings of the U.S. Bureau of the Census Annual Research Conference. Washington, D.C.: U.S. Census Bureau.

Zanutto, Elaine, and Zaslavsky, Alan M. (1996). Modeling census mailback questionnaires, administrative records, and sampled nonresponse followup, to impute census nonrespondents. In Proceedings, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Zanutto, Elaine, and Zaslavsky, Alan M. (2001). Using administrative records to impute for nonresponse. To appear in R. Groves, R.J.A. Little, and J.Eltinge (Eds), Survey Nonresponse. New York: John Wiley.

Zaslavsky, A., and Wolfgang, G. (1993). Triple system modeling of census, post-enumeration survey, and administrative-list data. Journal of Business and Economic Statistics, 11: 279-288.

Tables and Figures

Table 1: The Edited Person File Layout and Master Housing File Layout

The 204-character EPF standard output layout is presented below. This file also serves as the input **and** output layouts for SSN Validation and Search and the input layout for Person Processing. Person Processing uses this layout as the basis for creating the Composite Person Record. Although the EPF is output in SAS format, the record layout is presented in "flat file" format to facilitate descriptions of field contents. (Source: U.S. Census Bureau, StARS Person Edit Programming Specifications: StARS 1999 Development Person Edit General Specifications, Catalog Number: STAR9904-00.)

#	Field	Field Description	Length	Positions Beg - End	
1.	UID	Unique Identifier	17	1 - 17	CHAR
2.	SSN	Social Security Number (SSN)	9	18 - 26	CHAR
3.	SSNSRC	SSN Source (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	1	27 - 27	CHAR
4.	ARFNM	Administrative Record (ADREC) First Name	15	28 - 42	Char
5.	ARFNMO	ADREC First Name Overflow Flag 0 = No Character Overflow 1 = Additional Characters in Name Field	1	43 - 43	Char
6.	ARMNM	ADREC Middle Name	15	44 - 58	Char
7.	ARMNM2	ADREC Second Middle Name	15	59 - 73	Char
8.	ARLNM	ADREC Last Name	20	74 - 93	Char
9.	ARLNMO	ADREC Last Name Overflow Flag 0 = No Character Overflow 1 = Additional Characters in Name Field	1	94 - 94	Char
10.	ARSUFFIX	ADREX Name Suffix Generation Flag Equivalent (from the name standardizer) Generation Flag / Suffix 0 = Blank 1 = JR 2 = III 3 = IV 4 = SR	3	95 - 97	Char
11.	STFNM	Standardized First Name	15	.98 - 112	Char
12.	STMNM	Standardized Middle Name	15	113 - 127	Char
13.	STMNM2	Standardized Second Middle Name	15	128 - 142	Char
14.	STLNM	Standardized Last Name	20	143 - 162	Char

15.	STDECD	Standardized Deceased Flag (returned by standardizer) 0 = Not Deceased (default) 1 = Deceased	1	163	-	163 Char
16.	NAMESRC	Name Source 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	1	164	-	164 Char
17.	SEX	Gender Blank = No Data Present/NA 1 = Male 2 = Female	1	165	-	165 Char
18.	SEXSRC	Sex Source 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	1	166	-	166 Char
19.	CYDOB	Birth Date (Century and Year) Blank = No Data Present/NA CCYY format – valid range between 1880 – 1999 (inclusive)	4	167	-	170 Char
20.	MMDOB	Birth Date (Month) Blank = No Data Present/NA MM format – valid range between 01 – 12	2	171	-	172 Char
21.	DDDOB	Birth Date (day) Blank = No Data Present/NA DD format – valid range between 01 -31	2	173	-	174 Char
22.	DOBSRC	Date of Birth Source 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	1	175	-	175 Char

23. DOBSCOR — Date of Birth Score — 2 — 176 - 177 Char

* = non blank

YYYY	MM	DD	Score
*	*	*	14
*	*	-	12
*	-	*	10
*	-	-	08
-	*	*	06
-	*	-	04
-	-	*	02
-	-	-	00

24. PRIM — Record Primary Person Flag — 1 — 178 - 178 Char

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

25. HISP — Hispanic Origin — 1 — 179 - 179 Char

Blank = No Data Present/NA
1 = Hispanic
2 = Not Hispanic

26. HISPSRC — Hispanic Source — 1 — 180 - 180 Char

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

27. RACE — Race — 1 — 181 - 181 Char

Blank = No Data Present/NA
1 = White
2 = Black
3 = American Indian, Eskimo, or Aleut
4 = Asian or Pacific Islander
5 = Other
6 = Unknown

28. RACESRC — Race Source — 1 — 182 - 182 Char

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

29. DECSSRC — Deceased Source — 1 — 183 - 183 Char

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

30. DCSTYP — Deceased Indicator Type — 2 — 184 - 185 Char

* = indication person is deceased
AN = ADREC Name Standardization
AV = ADREC Deceased Variable
PC = PCF Mortality Model
NV = NUMIDENT Deceased Variable

AN	AV	PC	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

31. CYDOD — Date of Death (Century and Year) — 4 — 186 - 189 Char

Blank = No Data Present/NA
CCYY format – valid range
between 1880 – 1999 (inclusive)

32. MMDOD — Date of Death (Month) — 2 — 190 - 191 Char

Blank = No Data Present/NA
MM format – valid range
between 01 –12

33. DDDOD — Date of Death (day) — 2 — 192 - 193 Char

Blank = No Data Present/NA
DD format – valid range
between 01 -31

34. CYADCYDA — ADREC Cycle Date (Century and Year) — 4 — 194 - 197 Char
CCYY format – valid range between 1880 – 1999 (inclusive)

35. MMADCYDA — ADREC Cycle Date (Month) — 2 — 198 - 199 Char
Blank = No Data Present/NA
MM format – valid range between 01 –12

36. DDADCYDA — ADREC Cycle Date (Day) — 2 — 200 - 201 Char
Blank = No Data Present/NA
DD format – valid range between 01 -31

37. NMSCORE — Name Score — 2 — 202 - 203 Char
(Calculated from ADREC name)
F=Full I=Initial * = non blank

Las	Fir	Mid	J/S	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 score as follows:

Suffix = + 1
Middle Initial = + 2
Middle Full = + 6
First Initial = + 8
First Full = + 24
Last Name = + 32

38. VERFLG — SSN Verification Flag — 1 — 204 - 204 Char
Set to "0" (default) for EPF.

Master Housing File

The 375-character MHF record layout is provided below. One thousand 3-Digit ZIP Files comprise the MHF. Specific definitions for each type of HUID are contained within the field description area of the record layout. Although the MHF is maintained as a SAS data set, an ASCII format record layout is provided to facilitate field descriptions.

FIELD	FIELD DESCRIPTION	LENGTH
AID	POINTER TO ORIGINAL SOURCE RECORD UIDs	11
FIPST90	90 FIPS STATE CODE	2
FIPCTY90	90 FIPS COUNTY CODE	3
TRACT90	90 "TABULATION" CENSUS TRACT CODE char 1-4 = tract char 5-6 = tract suffix ('0' fill)	6
BLOCK90	90 "TABULATION" BLOCK (INCLUDING SUFFIX) char 1-3 = block char 4 = suffix reflecting 1990 split char 5 = suffix reflecting current split (since 1990)	5
CONFIDFG	CONFIDENCE FLAG FOR 2000 CODING 1 ZIPCODE AND STREET NAME EXACT MATCH TO TIGER 2 EQUIVALENT MATCH TO ZIP, PERFECT MATCH TO STREET ADDRESS 3 ZIPCODE EXACT MATCH AND EQUIVALENT MATCH TO STREET NAME 4 ZIPCODE DIDN'T MATCH BUT FOUND EQUIVALENT STREET NAME 6 NOT CODED, 5-DIGIT ZIP NOT IN TIGER 7 NOT CODED, 5-DIGIT ZIP IS IN TIGER, AND ADDRESS IS NOT CITY STYLE, RURAL ROUTE OR PO BOX 8 NOT CODED, 5-DIGIT ZIP IS IN TIGER AND ADDRESS IS RURAL ROUTE OR PO BOX 9 NOT CODED, 5-DIGIT ZIP IS IN TIGER AND ADDRESS IS CITY STYLE	1
FIPST2K	2000 FIPS STATE CODE (input value if no Tiger match)	2
FIPCTY2K	2000 FIPS COUNTY CODE (input value if no Tiger match)	3
ZIP532K	2000 3-DIGIT ZIP (first 3 digits of ZIP Code) (input value if no Tiger match)	3

ZIP522K	2000 ZIP CODE (last 2 digits of ZIP Code) (input value if no Tiger match)	2
ZIP42K	2000 ZIP+4 (input value if no Tiger match)	4
BLOCK2K	2000 "COLLECTION" BLOCK (blank if no Tiger match)	5
BLKSFX2K	2000 "COLLECTION" BLOCK SUFFIX (blank if no Tiger match)	1
CFO2K	2000 CENSUS FIELD OFFICE (blank if no Tiger match)	4
LCO2K	LOCAL CENSUS OFFICE (blank if no Tiger match)	4
AREXSITE	AREX TEST SITE FLAG ' ' Not in an AREX test site 'G' Geocoded to AREX test site state and county 'Z' Zip code is in AREX test site state and county but not geocoded	1
TIGERID	TIGERLINE ID	10
SIDEID	TIGERLINE ID SIDE	1
TEA	TYPE OF ENUMERATION AREA (ONLY FOR ADDRESSES WITH 2000 BLOCK) (blank if no Tiger match) 1 ADDRESS INSIDE OF THE BLUELINE 2 ADDRESS LISTING OUTSIDE OF BLUELINE	1
IADDRESS	INPUT ADDRESS	60
CD1CITY	CITY NAME (from the IMHF)	35
	STANDARDIZED INPUT ADDRESS FIELDS:	
STHNP	HOUSE NUMBER PREFIX	2
STHN	HOUSE NUMBER	8
STHNP2	SECONDARY HOUSE NUMBER PREFIX SEPARATOR	2
STHN2	SECONDARY HOUSE NUMBER	6
STHNSX	HOUSE NUMBER SUFFIX	3
STSTPXDR	STREET NAME PREFIX DIRECTION	2
STSTPXTY	STREET NAME PREFIX TYPE	4
STSTNAME	STREET NAME	28
STSTSXTY	STREET NAME SUFFIX TYPE	4
STSTSXDR	STREET NAME SUFFIX DIRECTION	2
STSTEXT	STREET NAME EXTENSION INDICATOR	3
STSTRDSC	STRUCTURE DESCRIPTION	4
STSTRID	STRUCTURE IDENTIFIER	6
STWSTDSC	WITHIN STRUCTURE DESCRIPTION	4
STWSTRID	WITHIN STRUCTURE IDENTIFIER	9
STRRDSC	RURAL ROUTE DESCRIPTION	4
STRRID	RURAL ROUTE IDENTIFIER	4
STBOXDSC	BOX DESCRIPTION	6
STBOXID	BOX OR IDENTIFIER	6
STPTPXDR	PROPERTY DESCRIPTION PREFIX DIRECTION	2
STPTPXTY	PROPERTY DESCRIPTION PREFIX TYPE	4
STPTNAME	PROPERTY DESCRIPTION NAME	28
STPTSXTY	PROPERTY DESCRIPTION SUFFIX TYPE	4
STPTSXDR	PROPERTY DESCRIPTION SUFFIX DIRECTION	2

ADDRTYPE — STANDARDIZED ADDRESS TYPE FLAG — 1

0 = No Input Address
1 = Non-Standardized Address
2 = Standardized Street Address
3 = Standardized P.O. Box Address
4 = Standardized Property Address
5 = Standardized Street & P.O. Box Address
6 = Standardized Street & Property Address
7 = Standardized Property & P.O. Box Address
8 = Rural Route (standardized with description or identifier)
9 = Undefined Address
A = Standardized Street and Rural Route Address
B = Standardized Property and Rural Route Address

GEOTYPE — TYPE OF GEOCODING — 2

TIGER	MAFID	BLK2K	BLK90	TYPE
-	-	-	-	00
-	-	-	*	01
-	-	*	-	02
-	-	*	*	03
-	*	-	-	04
-	*	-	*	05
-	*	*	-	06
-	*	*	*	07
*	-	-	-	08
*	-	-	*	09
*	-	*	-	10
*	-	*	*	11
*	*	-	-	12
*	*	-	*	13
*	*	*	-	14
*	*	*	*	15

HUID — HOUSING UNIT IDENTIFIER — 35

(First character defines address category)
(read: category – identifier construction sequence)

8 - ADDRESS WITH TIGER ID (applicable only if non-blank house number and Tiger ID)
"8"//FIPS State//FIPS County//Tiger Line ID//Side ID//House Number (6)//APT Number (6)//Structure ID (6)

7 – ADDRESS WITH 2000 BLOCK (applicable only if non-blank house number, no Tiger ID, & 2K Block)
"7"//FIPS State//FIPS County//"B"//"000"//2K Block//2K Block Suffix//"B"//House Number (6) Apt Number (6)//Structure ID (6)

6 – ADDRESS WITH MAFID (applicable only if non-blank house number, no Tiger, no 2K Block, and MAFID)
"6"//FIPS State//FIPS County//"N"//Sequence #//"N"//House Number (6)//Apt Number (6)//Structure ID (6)

Note: This HUID designator reserved for future use.

5 – OTHER ADDRESS (applicable only if non-blank house number, no Tiger ID, no 2K Block, and No MAFID)
"5"//FIPS State//FIPS County//"O"//Sequence #//"O"//House number (6)//Apt Number (6)//Structure ID (6)

4 – BOX ADDRESS
"4"//FIPS State//FIPS County//ZIP//Rural Route Description//Rural Route ID//Box Number//Blanks (9)

3 – GEOCODED PROPERTY ADDRESS
"3"//FIPS State//FIPS County//Tiger ID//Side ID//Property Name (1 to 18, blank fill)

2 – NON-GEOCODABLE ADDRESS
"2"//FIPS State//FIPS County//Blank fill (29)

1 – NON-STANDARDIZED ADDRESS
"1"//FIPS State//FIPS County//Blank fill (29)

0 – NO ADDRESS ON FILE
"0"//FIPS State//FIPS County//Blank fill (29)

ADDRSRC — ADDRESS SOURCE INDICATOR — 6

(After final unduplication – a value in the field may indicate a combined address source)
Values (0 or 1) indicate presence or absence of record on original source file(s) as follows:
Position:
1 = I (IRS 1040 File)
2 = J (IRS 1099 File)
3 = M (Medicare File)
4 = T (Indian Health Service)
5 = V (Selective Service)
6 = W (HUD TRACs file)

BSACOD — COMMERCIAL FLAG FROM ABI FILE — 1

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 and 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 Agents' Offices
9 = Mixed Use – Other Than Type 8
A = Unmatched to Commercial

BSASRC — BSA COURCE CODE — 3

Bit Set	**Meaning**	
	Residential – Known Single Unit	Not Applicable// reserved for future use.
	Residential – Possible Multi-Unit	
0	Apartment Buildings	
1	Hotels/Motels	
2	Mobile Home Parks/Marinas/RV Parks and 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 Agents' Offices	
7	Mixed Use – Other Than Type 8	
8	Unmatched to Commercial	

PRISIC — PRIMARY SIC CODE FROM ABI FILE — 6

(see attachment 1 to this specification)

ATHOME — WORK AT HOME FLAG — 1

"1" indicates a business at home

ABINUM	ABI RECORD REFERENCE NUMBER (blank if BSACOD = A) (the ABI unique identifier for each business in the database).	9

Table 2: Draft of the **StARS COMPOSITE PERSON RECORD LAYOUT**

Total Length : **314**
Date Created : **08/14/2000**
Date Revised : N/A

#	Field	Field Description	Length	Positions Beg - End	
		*** Record Information ***			
1.	SSN	Social Security Number (SSN) **Note:** "Dummy" SSN created for non-verified records and records with no SSN (invalid or blank)	9	1 - 9	Char
2.	ARSOURCE	Admin Record Source (Tally) Valid Value = 0 – 9 (9 = 9 or more occurrences) Pos. 1: = TY 98 IRS 1040 (1040) Pos. 2: = TY 98 IRS 1099 (1099) Pos. 3: = Medicare (MEDB) Pos. 4: = Blank Pos. 5: = Indian Health Service (IHS) Pos. 6: = Selective Service (SSS) Pos. 7: = HUD TRACs (TRAC)	7	10 - 16	Char
3.	SSNSRC	SSN Source Code Position: 1 2 3 4 5 6 7 1040 1099 MEDB CNUM IHS SSS TRAC Value Meaning (relative to selected value) 0 = Data not reported on record 1 = All records agree (within source) 2 = All records disagree (non-blank) 3 = Conflict within source file (2 + 1) 4 = Input record blank 5 = Within source agree and blank (4 + 1) 6 = Within source disagree and blank (4 + 2) 7 = Within source agree, disagree, & blank (4+2+1)	7	17 - 23	Char
		*** Address Information ***			

4.	HUID	Admin Record HUID
		*ADDRESS WITH TIGER ID =
		Category "**8**"//FIPST2K //FIPCTY2K //TIGERID //SIDEID //
		STHN(6) //SUBSTRUCTURE (12)
		(Applicable only if House Number & Tiger ID are not blank)
		*ADDRESS WITH 2000 BLOCK =
		Category "**7**"//FIPST2K //FIPCTY2K //"B" //"000" //BLOCK2K//
		BLKSFX2K //"B" //STHN(6) //SUBSTRUCTURE 12
		(Applicable only if House Number and 2K Block are not blank and Tiger ID is blank)
			35	24 - 58	Char

#	Field	Field Description	Length	Beg - End

*Address with MAF ID =
Category "**6**"//FIPST2K //FIPCTY2K //ZIP5 //Sequence # (6) //
STHN (6) //SUBSTRUCTURE (12)
(Applicable only if House Number and MAFID are not
blank and TIGER ID and 2K Block are blank)

*Other Address =
Category "**5**"//FIPST2K //FIPCTY2K //ZIP5 //Sequence # (6) //
STHN (6) //SUBSTRUCTURE (12)
(Applicable only if House Number is not blank
TIGER ID, 2K Block, and MAFID are blank)

*Box Address =
Category "**4**"//FIPST2K //FIPCTY2K //ZIP5 //STRRDSC (4)//
STRRID (4) //STBOXID (6) //Blanks(10)

*Geocoded Property Address =
Category "**3**"//FIPST2K //FIPCTY2K //TIGERID //SIDEID //
STPTNAME (18) - blank fill as required

*Non-Geocodable Address =
Category "**2**"//FIPST2K //FIPCTY2K //ZIP5 //Sequence # (12) //
Blank Fill (12)

*Non-Standardized Address =
Category "**1**"//FIPST2K //FIPCTY2K //Blank Fill (29)

*No Address on File =
Category "**0**"//FIPST2K //FIPCTY2K //Blank Fill (29)

5. HUIDSRC Admin Record Source of HUID 7 59 - 65 Char

Position:	1	2	3	4	5	6	7
	1040	1099	MEDB	CNUM	IHS	SSS	TRAC

Value Meaning (relative to selected value)
0 = Data not reported on record
1 = All records agree (within source)
2 = All records disagree (non-blank)
3 = Conflict within source file (2 + 1)
4 = Input record blank
5 = Within source agree and blank (4 + 1)
6 = Within source disagree and blank (4 + 2)
7 = Within source agree, disagree, & blank (4+2+1)

6. HUIDECFL Selected HUID Decision Flag 1 66 - 66 Char
Value indicates selection rule invoked:
1 = xx

Place holder for address selection rules

2 = xx

7. FIPST2K 2000 FIPS State Code 2 66 - 67 Char

8. FIPCTY2K 2000 FIPS County Code 3 68 - 70 Char

#	Field	Field Description	Length	Beg - End	
9.	ZIP532K	2000 ZIP Code (first 3 digits of 5-Digit ZIP Code)	3	71 - 73	Char
10.	ZIP522K	2000 ZIP Code (last 2 digits of 5-Digit ZIP Code)	2	74 - 75	Char
11.	ZIP42K	2000 ZIP Code + 4	4	76 - 79	Char
12.	BLOCK2K	2000 "Collection" Block	5	80 - 84	Char
13.	BLKSFX2K	2000 "Collection" Block Suffix	1	85 - 85	Char
14.	TRACT90	1990 Census Tabulation Tract Code Characters 1-4 = Tract 5-6 = Tract Suffix	6	86 - 91	Char
15.	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 and 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 Agents' Offices 9 = Mixed Use – Other Than Type 8 A = Unmatched to Commercial	1	92 - 92	Char
16.	ATHOME	Work At Home Flag (From ABI File) Blank = Default (not flagged as work at home) 1 = Identified as work at home	1	93 - 93	Char
17.	PROXYFLG	Proxy Flag	1	94 - 94	Char
		* **Name Information** *			
18.	ARFNM	Administrative Record (ADREC) First Name	15	95 - 109	Char
19.	ARFNMO	ADREC First Name Overflow Flag 0 = No Character Overflow 1 = Additional Characters in Name Field	1	110 - 110	Char
20.	ARMNM	ADREC Middle Name	15	111 - 125	Char
21.	ARMNM2	ADREC Second Middle Name	15	126 - 140	Char
22.	ARLNM	ADREC Last Name	20	141 - 160	Char
23.	ARLNMO	ADREC Last Name Overflow Flag 0 = No Character Overflow 1 = Additional Characters in Name Field	1	161 - 161	Char

#	Field	Field Description	Length	Beg	-	End	
24.	ARSUFFIX	ADREX Name Suffix Name Standardizer Generation Flag Equivalent Generation Flag = Suffix 0 = Blank 1 = JR 2 = III 3 = IV 4 = SR	3	162	-	164	Char
25.	STFNM	Standardized First Name	15	165	-	179	Char
26.	STMNM	Standardized Middle Name	15	180	-	194	Char
27.	STMNM2	Standardized Second Middle Name	15	195	-	209	Char
28.	STLNM	Standardized Last Name	20	210	-	229	Char
29.	NMSCORE	Name Score (Calculated from ADREC name) F=Full I=Initial * = non blank	2	230	-	231	Char

Las	Fir	Mid	J/S	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 score as follows:

Suffix	= + 1
Middle Initial	= + 2
Middle Full	= + 6
First Initial	= + 8
First Full	= + 24
Last Name	= + 32

#	Field	Field Description	Length	Beg	-	End	
30.	NAMESRC	Name Source	7	232	-	238	Char

Position: 1 2 3 4 5 6 7
1040 1099 MEDB CNUM IHS SSS TRAC

Value Meaning (relative to selected value)

0 = Data not reported on record
1 = All records agree (within source)
2 = All records disagree (non-blank)
3 = Conflict within source file (2 + 1)
4 = Input record blank
5 = Within source agree and blank (4 + 1)
6 = Within source disagree and blank (4 + 2)
7 = Within source agree, disagree, & blank (4+2+1)

31.	PRIM	Primary Person Record Flag	1	239	-	239	Char

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

32.	NMDECFL	Selected Name Decision Flag	1	240	-	240	Char

Value Indicates Selection Rule Invoked as follows:
1 = Most Recent (Last) Name
2 = Most Complete Name Matching Most Recent Last Name
3 = Most Recent
4 = Highest Name Score

*** ***
Standardized Person Data (Demographics)
*** ***

33.	SEX	Gender	1	241	-	241	Char

1 = Male
2 = Female

#	Field	Field Description	Length	Beg	-	End	
34.	SEXSRC	Gender Source	7	242	-	248	Char

Position: 1 2 3 4 5 6 7
1040 1099 MEDB CNUM IHS SSS TRAC

Value Meaning (relative to selected value)
0 = Data not reported on record
1 = All records agree (within source)
2 = All records disagree (non-blank)
3 = Conflict within source file (2 + 1)
4 = Input record blank
5 = Within source agree and blank (4 + 1)
6 = Within source disagree and blank (4 + 2)
7 = Within source agree, disagree, & blank (4+2+1)

35.	SEXDECFL	Selected Gender Decision Flag	1	249	-	249	Char

Selection Rule Invoked:
1 = Male if record appears on SSS
2 = Most frequent observation
3 = PCF probability

36.	RACE	Selected Race	1	250	-	250	Char

1 = White
2 = Black
3 = American Indian, Eskimo, or Aleut
4 = Asian or Pacific Islander
5 = Other
6 = Unknown

37.	RACESRC	Race Source	7	251	-	257	Char

Position: 1 2 3 4 5 6 7
1040 1099 MEDB CNUM IHS SSS TRAC

Value Meaning (relative to selected value)
0 = Data not reported on record
1 = All records agree (within source)
2 = All records disagree (non-blank)
3 = Conflict within source file (2 + 1)
4 = Input record blank
5 = Within source agree and blank (4 + 1)
6 = Within source disagree and blank (4 + 2)
7 = Within source agree, disagree, & blank (4+2+1)

38.	RACDECFL	Selected Race Decision Flag	1	258	-	258	Char

Selection Rule Invoked:
1 = AI/AN flag from IHS file
2 = Most frequent observation
3 = PCF Probability

39.	HISP	Hispanic Origin	1	259	-	259	Char

1 = Hispanic
2 = Not Hispanic

#	Field	Field Description	Length	Beg	-	End	
40.	HISPSRC	Hispanic Source	7	260	-	266	Char

Position: 1 2 3 4 5 6 7
1040 1099 MEDB CNUM IHS SSS TRAC

Value Meaning (relative to selected value)

0 = Data not reported on record
1 = All records agree (within source)
2 = All records disagree (non-blank)
3 = Conflict within source file (2 + 1)
4 = Input record blank
5 = Within source agree and blank (4 + 1)
6 = Within source disagree and blank (4 + 2)
7 = Within source agree, disagree, & blank (4+2+1)

41. HISDECFL — Selected Hispanic Decision Flag — 1 — 267 - 267 Char

Selection Rule Invoked:
0 = No reported value
1 = Hispanic – most frequent observation
2 = Not Hispanic – most frequent observation
3 = PCF Value - Hispanic
4 = PCF Value – Not Hispanic

NOTE: PCF used only to resolve conflict between values 1 & 2 (blank values not considered)

42. CYDOB — Birth Date (Century and Year) — 4 — 268 - 271 Char

Blank = No Data Present/NA
CCYY format – Valid Range between 1880 – 1999

43. MMDOB — Birth Date (Month) — 2 — 272 - 273 Char

Blank = No Data Present/NA
MM format – Valid Range between 01 - 12

44. DDDOB — Birth Date (Day) — 2 — 274 - 275 Char

Blank = No Data Present/NA
MM format – Valid Range between 01 – 31

45. DOBSCOR — Date of Birth Score — 2 — 276 - 277 Char

*=non blank

CCYY	MM	DD	Score
*	*	*	14
*	*	-	12
*	-	*	10
*	-	-	08
-	*	*	06
-	*	-	04
-	-	*	02
-	-	-	00

#	Field	Field Description	Length	Beg	-	End	
46.	DOBSRC	Date of Birth Source	7	278	-	284	Char

Position: 1 2 3 4 5 6 7
1040 1099 MEDB CNUM IHS SSS TRAC

Value Meaning (relative to selected value)

0 = Data not reported on record
1 = All records agree (within source)
2 = All records disagree (non-blank)
3 = Conflict within source file (2 + 1)
4 = Input record blank
5 = Within source agree and blank (4 + 1)
6 = Within source disagree and blank (4 + 2)
7 = Within source agree, disagree, & blank (4+2+1)

47.	DOBDECFL	Selected DOB Decision Flag	1	285	-	285	Char

Selection Rule Invoked:

Rule	Flag Value
(1)	1 = Most Frequent Occurrence
(2)	2 = Highest DOB Score
(3)	Most Reliable Record – Priority as follows:
	3 = Medicare File
	4 = Selective Service File
	5 = Census NUMIDENT File
	6 = HUD TRACs File
	7 = Indian Health Services File
	Note: Flag values 3-7 indicate selection Rule 3
(4)	8 = Most Recent DOB
(5)	9 = First Record Read-in Among Ties

48.	CYDOD	Date of Death (Century and Year)	4	286	-	289	Char

Blank = No Data Present/NA
CCYY format – Valid Range between 1880 - 1999

49.	MMDOD	Date of Death (Month)	2	290	-	291	Char

Blank = No Data Present/NA
MM format – Valid Range between 01 – 12

50.	DDDOD	Date of Death (Day)	2	292	-	293	Char

Blank = No Data Present/NA
DD format – Valid Range between 01 – 31

51.	DODSRC	Date of Death Source	7	294	-	300	Char

Position: 1 2 3 4 5 6 7
1040 1099 MEDB CNUM IHS SSS TRAC

Value Meaning (relative to selected value)

0 = Data not reported on record
1 = All records agree (within source)
2 = All records disagree (non-blank)
3 = Conflict within source file (2 + 1)
4 = Input record blank
5 = Within source agree and blank (4 + 1)
6 = Within source disagree and blank (4 + 2)
7 = Within source agree, disagree, & blank (4+2+1)

#	Field	Field Description	Length	Beg	-	End	
52.	DODDECFL	Selected DOD Decision Flag Selection Rule Invoked: 1 = Most Complete DOD (implies most accurate) 2 = Record from Medicare and not NUMIDENT 3 = Most recent Medicare record	1	301	-	301	Char
53.	DCSTYP	Deceased Indicator Type * = indication person is deceased AN = ADREC Name Standardization AV = ADREC Deceased Variable PC = PCF Mortality Model NV = NUMIDENT Deceased Variable **Note:** Flag value represents a composite value from all sources based on recomputed bit tallies	2	302	-	303	Char

AN	AV	PC	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

54.	VERTYPE	SSN Verification Type Flag (from the SSN S&V program) Bit Set = Meaning 0 = SSN Not Matched or Invalid 1 = SSN Verified via name match criteria 2 = IRS 1040 Record – special match criteria 3 = IRS 1099 Record – special match criteria 4 = SSN Verified against "New SSN" List only (added to SSN Verified pool) 5 = SSN found in Search Pass 1, 2, or 3.	2	304	-	305	Char
55.	PIK	Protected Identification Key	9	306	-	314	Char

Table 3: Distribution of Race and Hispanic Origin in Proposed AREX 2000 Sites and the United States

	Baltimore* County, MD	Baltimore+ City, MD	Douglas* County, CO	El Paso+ County, CO	Jefferson* County, CO	United States
White	84.94%	39.09%	97.17%	85.99%	94.55%	80.29%
Black	12.35%	59.21%	0.67%	7.20%	0.74%	12.06%
American Indian, Eskimo, or Aleut	0.21%	0.35%	0.44%	0.82%	0.55%	0.79%
Asian or Pacific Islander	2.24%	1.08%	0.84%	2.48%	1.74%	2.92%
Other Race	0.26%	0.27%	0.87%	3.51%	2.42%	3.94%
Hispanic	1.17%	1.03%	3.16%	8.68%	7.02%	8.99%
Total Population	692,134	736,014	60,391	397,014	438,430	248,709,873
Total Housing Units	281,553	303,706	22,291	165,056	178,611	102,263,678

Source: American Factfinder 1990 Census Data

* Easy to Count Area
+ Hard to Count Area

Table 4: Numbers of records passing through each phase of Social Security Number Validation and Search, and their outcomes.

Total Records	
Number records to verification	874,715,694
Number records with NO SSN - no verification	1,035,279
Total number records	875,750,973
Verification	
Number records to verification	874,715,694
Number records verified	843,604,017
Number records not verified - to search	27,944,282
Number records not verified - NO search	3,167,395
Search	
Number records with NO SSN - to search	1,035,279
Number records not verified - to search	27,944,282
Total to search	28,979,561
Number records found	1,286,086
Number records CODE 9	55,193
Number records NOT found	27,638,282
Verified File	
Number records verified	843,604,017
Number records found	1,286,086
Number records CODE 9	55,193
Total Verified records	844,945,296
--> Number valid	
UnVerified File	
Number records not verified - NO search	3,167,395
Number records NOT found	27,638,282
Total Not-Verified records	30,805,677
--> Number invalid	

Summary		
Original Count	875,750,973	
Number valid	844,945,296	(96.5%)
Number invalid	30,805,677	(3.5%)
Final Count	875,750,973	

Notes: The Census Numident is regularly updated with a "new SSN" list as new SSNs are created and assigned. CODE9 refers to SSNs that are validated by being on this list. With the exception of CODE9's, "Validation" refers to a multi-stage attempt to match name/SSN/date of birth combinations to the Census Numident. "Search" refers to an attempt to use probabilistic matching methods to determine SSNs from name or name/date of birth information alone. 875,750,973 unduplicated person records from the StARS database were passed through this validation and search algorithm.

Table 5: Illustration of Matching Uncertainties In a Match between the Master Address File (MAF) and the CHUMS-Enhanced Internal Master Housing (IMH) File

CHUMS-enhanced IMH File						MAF				
A		Banana	St			1	Apple	St		
B	17	Banana	St			3	Apple	St	Apt	1
C	19	Banana	St	Apt	5	3	Apple	St	Apt	2
D	44	MLK, Jr.	Blvd			3	Apple	St	Apt	3
E	100	Route 4				3	Apple	St	Apt	4
F	7	Marie	Ln			7	Apple	St		
G		Wife Mrs. Smith				9	Apple	St		
H	5	Apple	St			#	Apple	St		
I	27	Apple	St			#	Martin Luther King, Jr.	Blvd		
J		Apple	St			#	Pennsylvania	Ave		
K	9999	Apple	St			7	Maria	Ln		
L	3	Apple	St	Apt	5					
M	1	Apple	St							
N	3	Apple	St	Apt	A					
O	3	Apple	St		ZZ					
P	3	Apple	St							
Q	3	Apple	St	Apt	1					

Notes: For the CHUMS-enhanced IMH file, column 1 labels the rows; column 2 is a house number; column 3 is a street name, column 4 is a street type; column 5 is a unit type identifier, and column 6 is a unit number. For the MAF, column 1 is a house number; column 2 is a street name; column 3 is a street suffix; column 4 is a unit type identifier; and column 5 is a unit number. All references are fictitious.

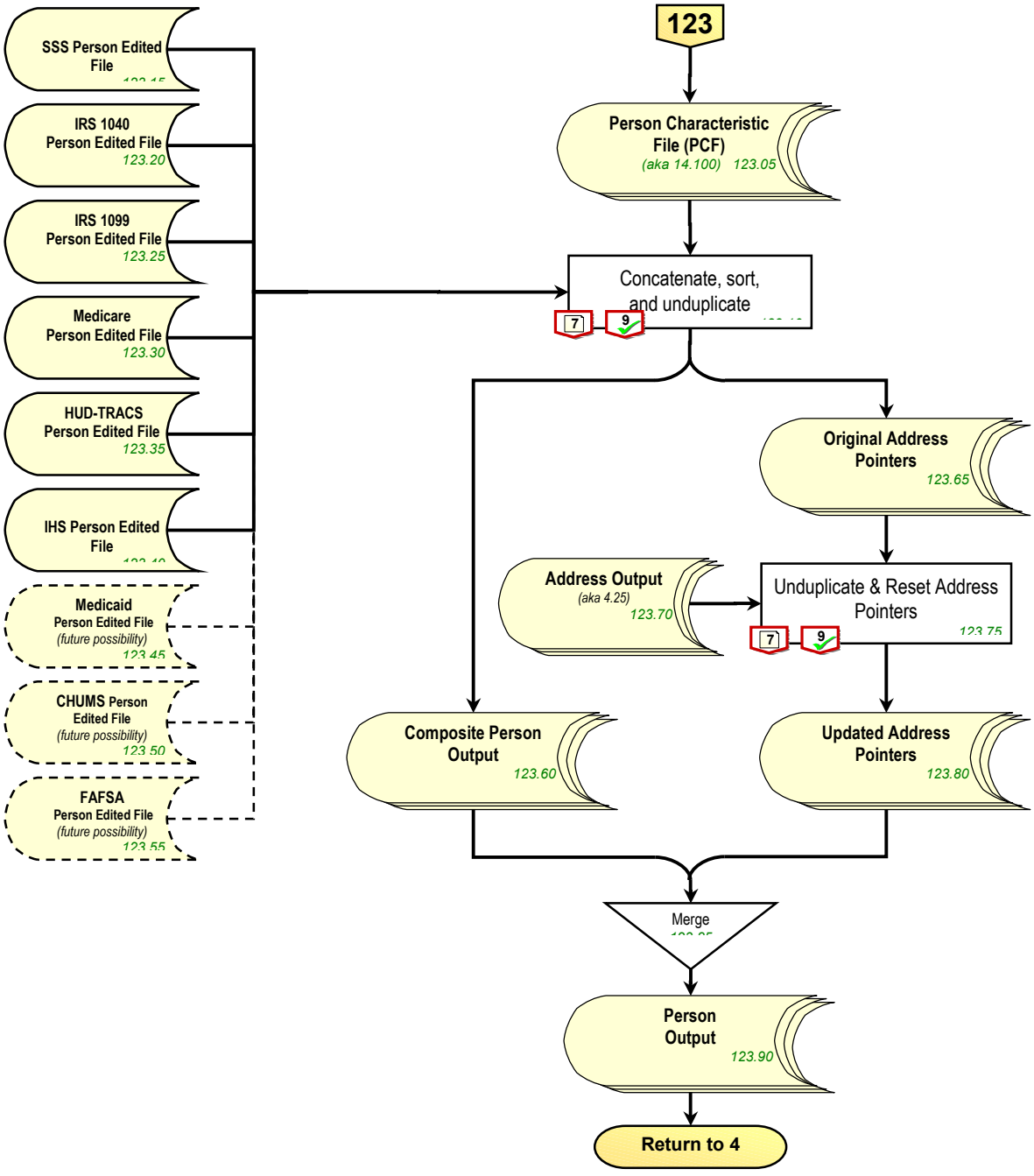
Table 6: Illustration of Potential Outcomes and Interpretations of these Outcomes in the IMH to MAF Match

Outcome of "CHUMS-enhanced IMH File" / MAF Match

MATCH				
Street	**BSA**	**BSA+Unit**	**Possible Explanations**	**Example**
NO	N/A	N/A	1 Street is not in MAF, either it was just missing or it's a new street	A,B,C
			2 Different, but valid representation of street name	D,E
			3 Misspelling of street name	F
			4 Erroneous street name	G
YES	NO	N/A	1 BSA is not in MAF, either it was just missing or it's a new BSA - There is a "hole" in MAF	H
			2 BSA is not in MAF, either it was just missing or it's a new BSA - A missing "street extension"	I
			3 Existing street with no incoming street number	J
			4 Erroneous street number	K
YES	YES	NO	1 Unit not in MAF, either it was just missing or it's a new unit	L
			2 Valid match - a BSA without separate units	M
			3 Different representation of a unit	N
			4 Erroneous unit information	O
			5 Missing unit information	P
YES	YES	YES	1 Valid match	Q

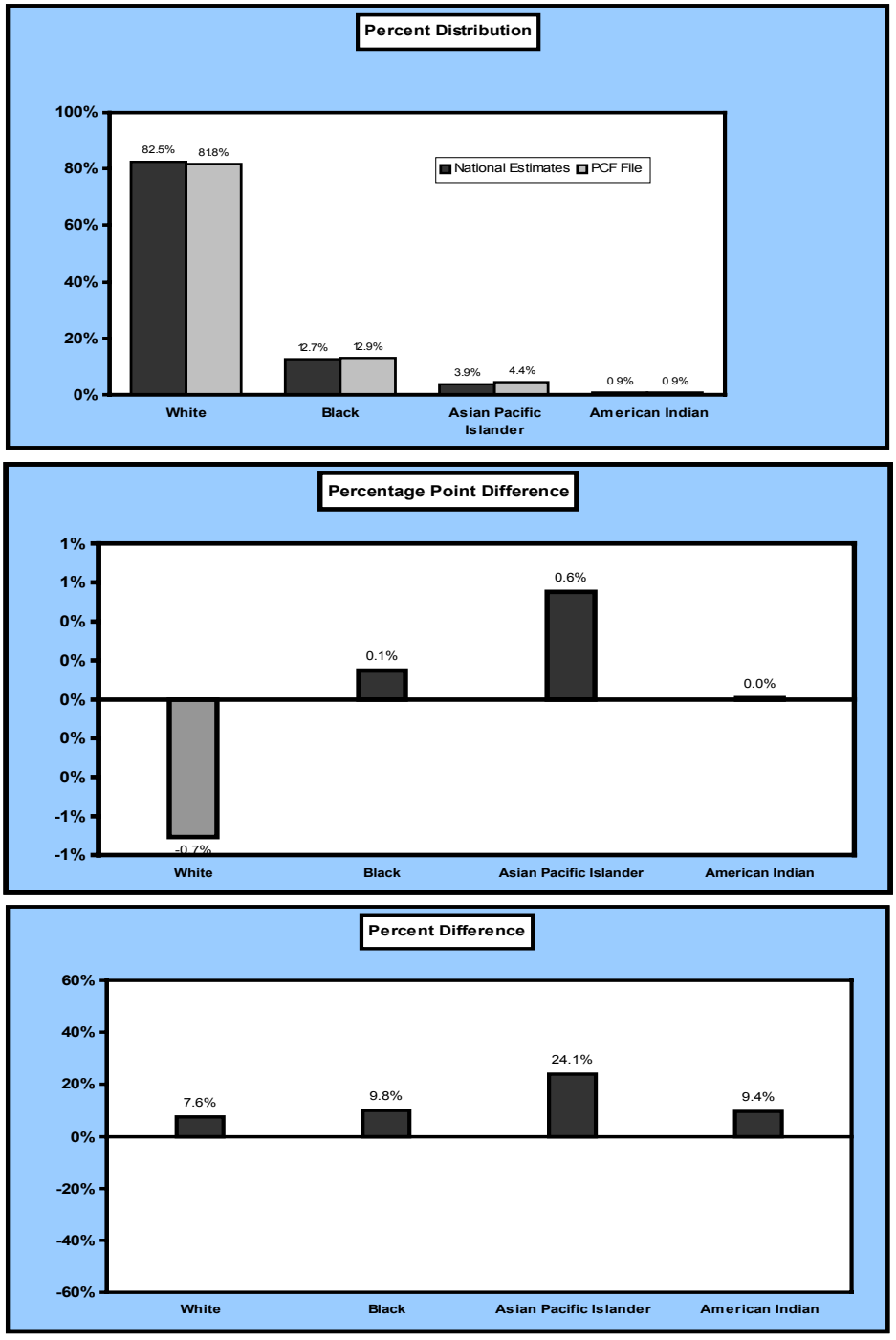
Notes: "Match" refers to the components of the address that match, where Street refers to street name, BSA refers to Basic Street Address (street number and name), and BSA+Unit refers to Basic Street Address plus, additionally, unit information match. "Possible explanations" provides a thumbnail of reasons why this combination might occur; and "Example" refers to the row of Table 5 corresponding to this situation. All examples are fictitious.

Figure 1: A diagrammatic depiction of files used to create the final StARS database[21]



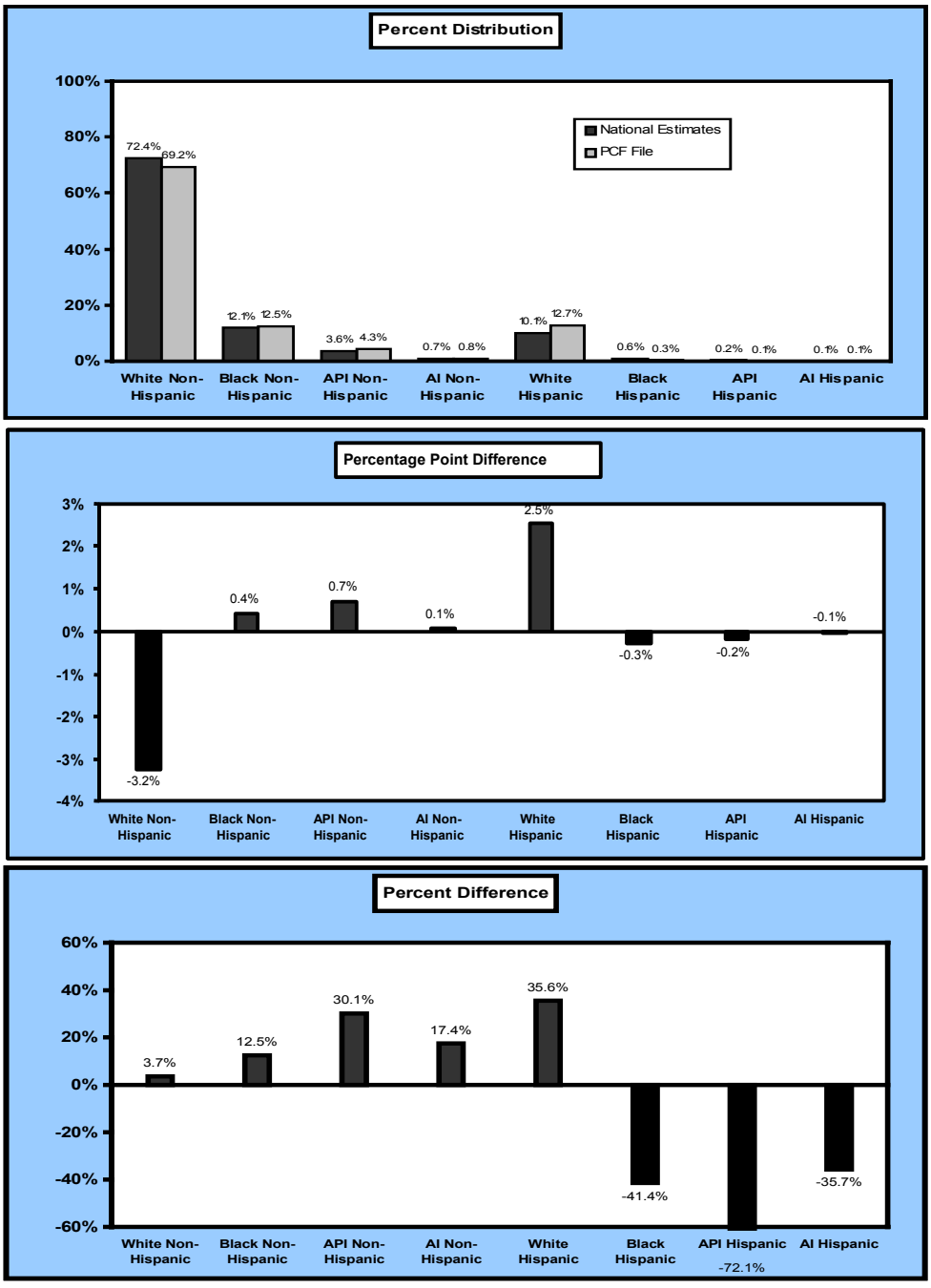
[21] Extracted from 1999 StARS Development and ARR File Management Flowcharts, U.S. Census Bureau, 2000.

Figure 2: Numeric and Percentage Point Differerences Between the National Estimates and the Personal Characteristics File (PCF) by Race: 1998



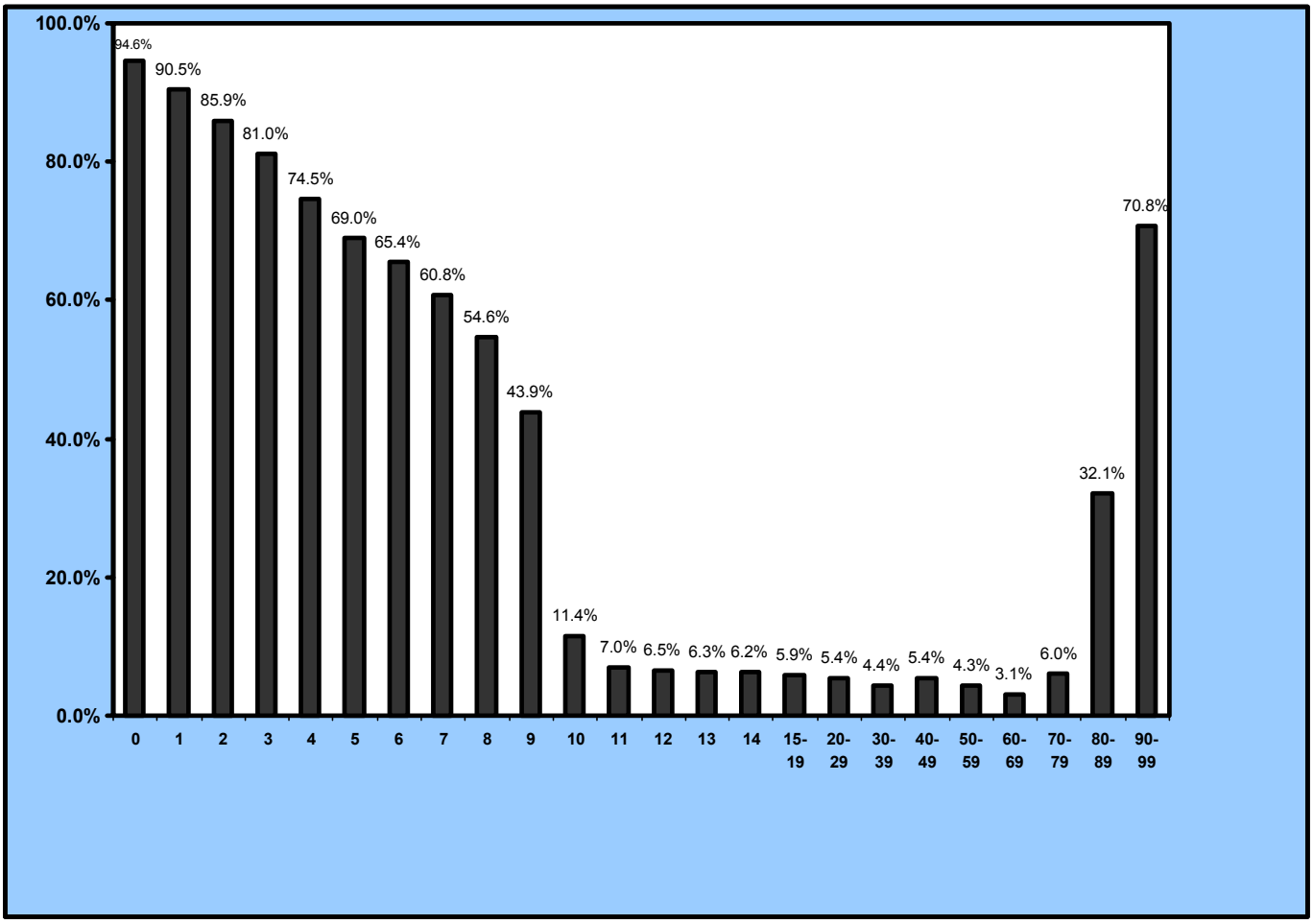
Note: "National" refers to 1998 national estimates; "PCF" refers to counts of modeled race in the Person Characteristics File.

Figure 3: Numeric and Percentage Point Differerences Between the National Estimates and the Personal Characteristics File (PCF) by Race and Ethnicity: 1998



Note: "National Estimates" refers to 1998 national estimates; "PCF Estimates" refers to counts of modeled race in the Person Characteristics File; WNH is white non-Hispanic; BNH is black non-Hispanic; APINH is asian/pacific islander non-Hispanic; AINH is american indian non-hispanic; WH is white Hispanic; BH is black Hispanic; APIH is asian/pacific islander Hispanic; and AIH is american indian Hispanic.

Figure 4: Distribution of Percent Missing for Original Race on the Numident File by Age: 1998



Note: Percentages are the percent within the age category that do not have an "original race" code on the Census Numident. This indicates that they have either not reported or been enumerated at birth. The increase at the end of the age distribution is due to the value of "not keyed". A record where the original race was "not keyed" was due to the computerization of the SSA records in 1970. Prior to 1970, if a person made a claim, any old paper records were thrown away and the original race was not keyed.

Figure 5: How Administrative Records are Created and Used

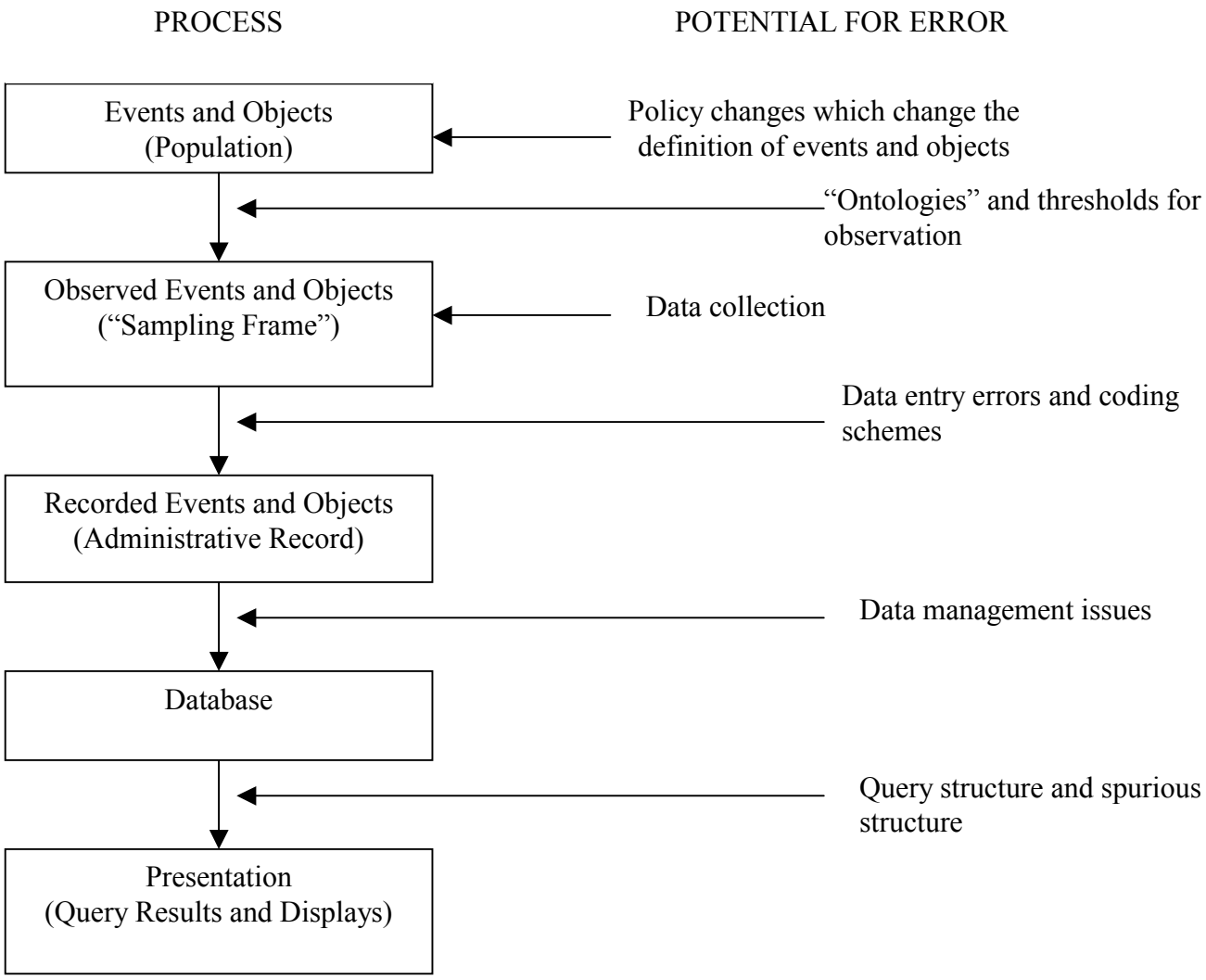
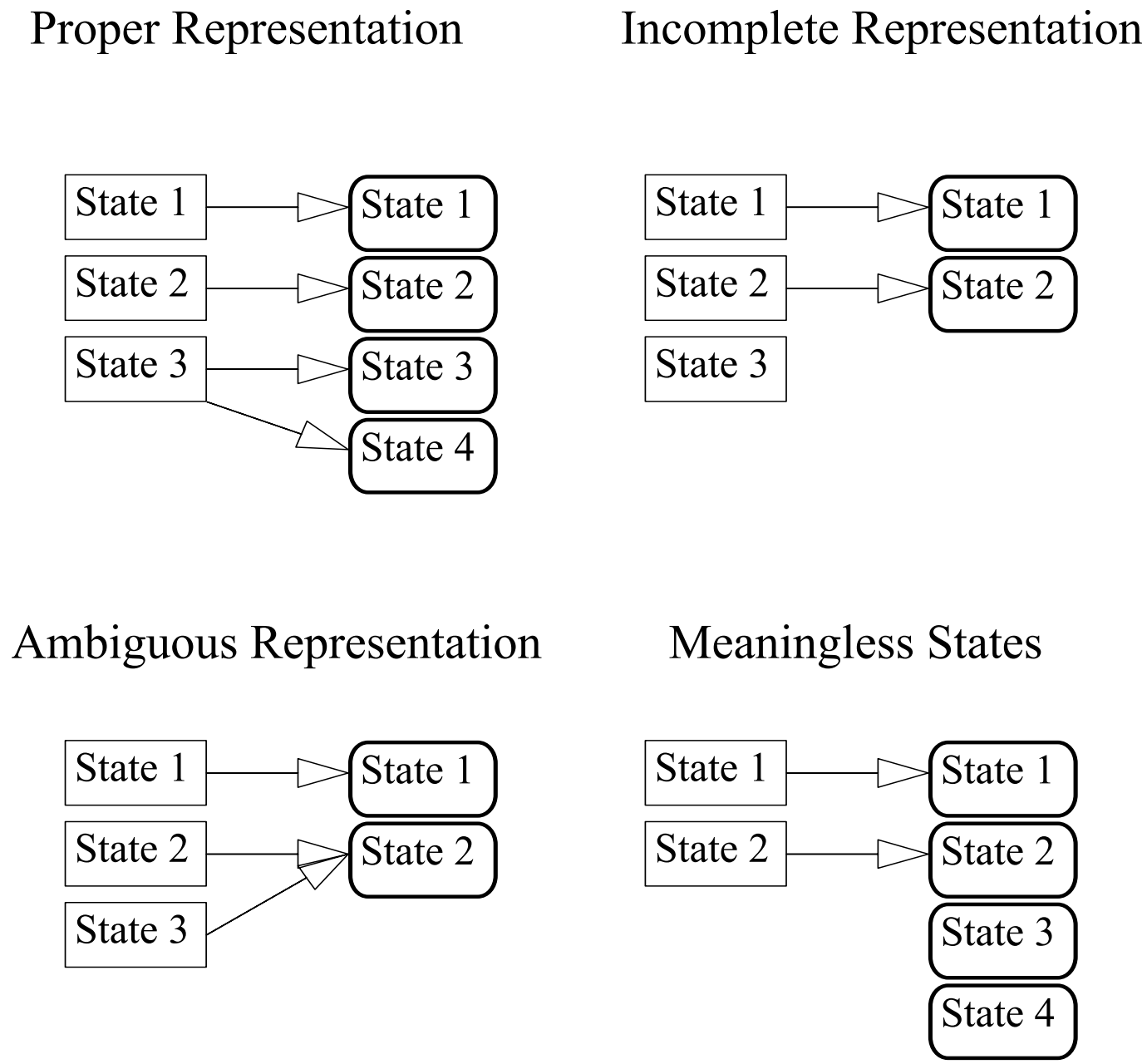


Figure 6: Four diagrams representing the various ways in which "real" data can be mapped into computer representation.



Source: Wand and Wang, 1996:90