

## ***Design Alternatives for Building Block Estimates***

by

Ron Prevost

Modeling and Applications Manager,  
Administrative Records Research Staff,  
Planning Research and Evaluation Division,  
Methodology and Standards Directorate,  
U. S. Bureau of the Census

### **Abstract**

The Census Bureau has been involved in a major research effort directed towards the expansion of administrative records utilization since 1996. We have begun the prototyping of a Statistical Administrative Records System (StARS) to combine several major federal administrative records systems such as the IRS individual tax returns and information returns, HCFA Medicare enrollment data, SSA Numident information, HUD Tract Rental Assistance Certification System, Indian Health Service, and Selective Service files. We have been researching several other federal micro and aggregate data files for Census Bureau statistical uses. This paper provides background information on StARS and proposes design alternatives for the creation of building block (census block and tract) population and housing estimates after 2000.

### **I. Why Research Administrative Records & What is StARS?**

The environment in which the Census Bureau must carry out its mission is changing dramatically. Costs for traditional data collection methods are increasing at the same time that federal budgets are shrinking and public cooperation is declining. Concurrently, the Census Bureau is faced with increasing demands for high quality statistics that are more current and that provide information for small geographic areas. Administrative records offer a solution for generating timely statistics at much lower costs while reducing respondent burden. Profound advances in computing technology and record linkage accuracy have significantly increased the feasibility of expanded uses of administrative records for statistical purposes that avoids repetitive and burdensome inquiries of the public. These advances coupled with spiraling costs of, and public resistance to, traditional data collection have increased the opportunity for significant benefits through using administrative records in data collection, estimation, and evaluation systems.

The Statistical Administrative Records System, or StARS, is a prototype research data system that employs administrative records from government agencies. The goal of this integrated and automated prototype system is to use information from administrative records to provide statistical and geographically referenced counts and characteristics of the population and housing of the United States. A fully developed and validated system supports the Census Bureau's goals for reducing respondent burden and costs. In addition, it facilitates the development of improved and/or new statistical products. If validation of the initial StARS prototype demonstrates that further efforts are required, research and analysis of the administrative records files comprising the model will still provide benefits to the Census Bureau in a less integrated approach.

## II. What Files are Components of our Administrative Records Research and the StARS Prototype?

The initial goals of our research are to provide new and improved addresses for post-2000 updates of the MAF, and to build a national portrait of housing units, households, and population by age, sex, and race, by assembling national files. Our first iteration of StARS will contain file systems that provide as broad and unbiased picture of our nation as possible. The following files are included in our research:

- IRS's Individual Master File-1040 Returns (117 million records)
- IRS's Information Returns Master File (750 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)
- American Business Information's National Business List (11 million records)

We are exploring the expansion of our file research for the year 2000 production cycle to contain data from the following national administrative records systems:

- 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 (*Student loan application databases*)

In order to accomplish this task we have purchased a powerful computer system with a large data storage capacity. The Census Bureau has also developed a strict data access policy to ensure the confidentiality of sensitive data by limiting access to a small handful of computer programmers and analysts.

The Administrative Records Research Staff (ARRS) has been conducting research on address sanitization and standardization. This research along with social security number and person name validation research is required to accurately unduplicate records on these large data sets. ARRS has been employing Group 1 Code 1 software to assist in our address sanitizing operations. Early tests of Code 1 show that it can process and sanitize about 95% of the addresses we find on our administrative records. We have been reviewing the records that did not match to Code 1's CASS compliant databases and we believe that we can further enhance address cleanup and matching through the use of probabilistic matching software. This Fall ARRS is conducting a national test of these operations.

We do not expect our first attempt at building this research system to provide unbiased results. The ARRS is working on an experiment in the year 2000 to conduct an Administrative Records Census Experiment, called AREX2000. This experiment will be operationalized in two sites within two states, Colorado and Maryland. These states were chosen because they contain counties where we expect administrative records to perform well (Baltimore County, MD and Douglas and Jefferson Counties CO), that are co-located with counties where we anticipate administrative records might not perform as well (Baltimore City, MD and El Paso County, CO). This selection allows us to perform field tests of bias on approximately 1 million housing units for the lowest possible cost. For these test sites, we hope to supplement our national administrative records' databases, with state Medicaid Enrollment records to eliminate a suspected bias of our current data files: persons in poverty. We will evaluate if state Medicaid data add coverage to our national files.

### III. Potential Customers of Administrative Records Research & StARS

The immediate users of products from a centralized administrative records effort are the Census Bureau's Demographic and Economic programs, with longer term benefits likely for the 2010 decennial census. Some of the program benefits will include:

- *Population Estimates:* An expansion of the intercensal estimates program using administrative records can produce timely, small area estimates of population totals and characteristics. Research towards the development of small area estimates (census blocks and tracts) will provide the ability to construct intercensal tabulations for Congressional Districts, urbanized areas, school districts, and other user defined areas. The expansion of administrative records capabilities also supports the Small Area Income and Poverty Estimates program.
- *Demographic Surveys:* In addition to improved survey controls, an expanded use of administrative records will result in improved survey estimates of population and housing characteristics. Administrative records offer a potential means for evaluating the quality

of survey responses and providing additional information for sample selection or stratification.

- *Economic Programs:* Emerging interest in linking of employer-employee data sets for economic and policy analysis is supported by a centralized administrative record research effort.
- *Master Address File/TIGER System:* Addresses from administrative records can be used cost effectively as a direct source of updates and as a means to target areas for improvement of the Master Address File.
- *Decennial Censuses:* Administrative records can be used to target special methods, improve coverage, and enhance imputation for missing responses. If developed to the fullest, and with broad acceptance, administrative records could ultimately become a primary data source for the headcount in 2010 with huge cost savings.

The long term beneficiaries of administrative records applications to official statistics are the public (as taxpayers and respondents), other government agencies (who use Census Bureau data), the Congress, and data users looking for timely, relevant statistics at reasonable costs.

#### IV. How Can Administrative Records Research & StARS Assist the Intercensal Estimates Program?

By exploiting administrative records more heavily in intercensal estimates and other demographic programs early in the decade, the Census Bureau will gain the experience needed to assess their potential for the decennial census in 2010. In 1994, the National Academy of Sciences Panel to Evaluate Alternative Census Methods said: “The future holds attractive prospects for using administrative records as the keystone in developing a greatly improved small-area demographic data system that can provide data more frequently at no increase and possibly a significant reduction in costs over the decade. However, these prospects can only be realized if the Census Bureau.....adopts a proactive policy to explore expanded uses of administrative records, with such policy to include.....a suitable organizational unit and adequate resources for research and development activities not tied directly to ongoing census and survey programs.” Two examples of potential StARS applications that could improve estimates are listed below.

##### *Developing New Products: Demographic Program Application*

The first example relates to the expansion of intercensal estimates to provide annual population estimates for small geographic areas. IRS individual tax returns (1040) and information returns (W-2 and 1099) files contain name, SSN, and current address information with population coverage estimated at 95% to 97% of the U.S. population (Mathematica Policy Research, Inc.,

1997). These files can be matched to the Social Security Administration's Numident file containing basic demographic information (age, gender, race/ethnicity) for each SSN holder. If addresses from the resulting matched file are linked to the Master Address File/TIGER System, tabulations for small geographic areas can be made on an annual basis. Research to develop the methodology, expand the coverage, and evaluate the quality of resulting tabulations is required. This process can create new products such as measures of household income and poverty status, by gender, race, and ethnicity of the householder as well as extending the estimation units below governmental unit geography to include census tracts and blocks.

*Reducing Cost and Improving Coverage: Address List Application*

The second example is an application that capitalizes on administrative records information to target improvements to the MAF/TIGER System. The major costs for address list building for Census 2000 include the full block canvass for areas with mostly house number and street name addresses and an address listing operation for remaining areas. Targeting of address list updating operations throughout the decade using American Community Survey community liaisons and administrative records has the potential to reduce the cost for these operations and improve the coverage of surveys and censuses that rely on the MAF/TIGER System. A full evaluation of these efforts is planned for 2005 to determine the extent to which canvassing operations for the 2010 Census might be eliminated or reduced. Exploratory research indicates that, in addition to targeting updates to the MAF/TIGER, addresses from administrative records files also can provide a cost effective mechanism for validating addresses supplied by other sources.

## V. What Are Building Block Estimates?

Building block estimates are estimates of the population, housing units, and their characteristics that can be attributed to Census blocks. These blocks can be aggregated to provide statistics for local governmental units, census tracts, school districts, congressional districts, and user defined areas. The integration of administrative records with field operations such as the Community Address Updating System, to continuously update the Master Address File will provide the address base necessary to develop current accurate estimates of housing stock for all blocks in the United States.

A variety of methods could be employed to create current annual estimates of the population for the United States and its subareas. The key to using any administrative records system for the development of population estimates is proper geographical coding. After solving geographic coding issues we can begin discussing the impact that individual record systems have on the methodology for estimating population, housing, and their characteristics. The following two pages present a continuum of design options for enhancing the current Intercensal Estimates Program.

### **Possibilities for Enhancement of Intercensal Estimation Operations**

*The purpose of this presentation is to explore thought-provoking ideas NOT planned activities.*

The following procedures assume that current methods are sound and can be enhanced by improved input data and processing techniques. Techniques that were applied during the 1990's include:

- State and county age, race, and gender estimates – Cohort Component
- State and county population estimate totals - Tax Method
- Local governmental unit estimates – Distributive Housing Method

### **How might data from StARS enhance the Intercensal Population Estimates Program?**

#### **Method 1 Improve IRS 1040 Processing**

- Merge with business data, flag and remove non-household records
- Combine data from multiple 1040 returns
- Merge to MAF/TIGER and maintain ID

#### **Method 2 Include IRS Information Returns**

- Treat all IRS information and 1040 returns as one system process like 1040's (Method 1)
- Merge to the MAF/TIGER and maintain ID

#### **Method 3 Improve Medicare Processing**

- Merge address information to MAF/Tiger and maintain ID
- Complete microdata edits and develop tabulations of total population by age and gender enrolled in Medicare.

#### **Method 4 Include Numident Data**

- Employ the Census numident file with micro-record modeling techniques pioneered by Barry Bye to assign age, race, ethnicity and gender characteristics to IRS 1040 and Information Returns system.
- Employ the Census numident file with micro-record modeling techniques pioneered by Barry Bye to assign age, race, ethnicity and gender characteristics to Medicare enrollment data.

#### **Method 5 Include StARS/CAUS Addresses**

- Employ all records processed by the StARS system to enhance the Master Address File. These records combined with field updates (CAUS system) will provide an up-to-date MAF.
- Note: This process suggests a change in the current housing unit estimation system by replacing building permit data with administrative records and field-based adds and deletes.

#### **Method 6 Include All StARS Components**

- Employ all records processed by the StARS system to develop estimates of families and households (an annual administrative records census without field operations).
- Note: This process suggests a major methodological change where population estimates might employ the StARS data as inputs. Efforts are then required to assess coverage and develop model-based coverage adjustment factors for population and housing characteristics.

Note: Being a design continuum, all enhancement methods build on their predecessors.

<b>Possible Post-2000 Estimates Method Designs - Change Continuum</b>						
	<b>Method 1</b>	<b>Method 2</b>	<b>Method 3</b>	<b>Method 4</b>	<b>Method 5</b>	<b>Method 6</b>
<b>1990's Base Product</b>	<b>Improvement over base</b>	<b>Improvement over previous method</b>				
State & County Population Characteristics	Better household income tallies (SAIPE)		Improved age & gender estimates for population aged 65 and older	Improved age, race, ethnicity & gender estimates for all cohorts		Improved age, race, ethnicity & gender estimates for all cohorts
State & County Population, Housing and Household Estimates	Better migration estimates for population under 65 years of age  Better household estimates	Universe improvement for IRS1040 non-filers.  Better migration and household estimates	Better migration estimates for population aged 65 and older.	Improved migration estimates for the population <65 and the population 65 and older  Improved state household estimates by age of head	Improved housing unit estimates  Improved alternative county and state population estimates based on the Housing Unit Method	Alternative and improved estimates for all system components
Local Governmental Unit Population & Housing Estimates	Better persons per household estimates	Large families accommodated – better persons per household estimates	Better county controls provide improved population estimates	Better county controls provide improved population estimates	Improved housing estimates result in improved local area estimates of population	Alternative and improved estimates for all system components
Building Block (Census Tract and block) Estimates of Population & Housing	Ability to create Distributive Housing Unit Method estimates based on address data and household size  Ability to create estimates for Congressional Districts, School Districts, and user defined areas	Improved coverage may be very area specific  Higher quality Distributive Housing Unit Method estimates	Better county controls provide improved population estimates	Better county controls provide improved population estimates	Vastly improved housing unit estimates are employed to improve population estimates	Alternative and improved estimates for all system components