METHODS FOR CONDUCTING AN ADMINISTRATIVE RECORDS EXPERIMENT IN CENSUS 2000

Charlene Leggieri, Arona Pistiner and James Farber, U.S. Census Bureau[1]
Arona Pistiner, U.S. Census Bureau, 4700 Silver Hill Road, Stop 9200, Washington, DC 20233, USA

Keywords: Statistical Administrative Records System

1. Introduction

The Administrative Records Census Experiment (AREX) 2000 was designed to provide information on the feasibility of conducting an administrative records census (ARC). An ARC is a census in which housing and population data are drawn from administrative records from various government agencies. The plan for Census 2000 explicitly called for experimentation with an ARC for two reasons. First, the use of administrative records as the primary data collection method has enormous potential for reduction in cost and response burden. Second, significant testing of administrative records was not done as part of the 1990 Census. As a result, the Census Bureau was not sufficiently prepared to consider using administrative records in Census 2000.

The administrative records used in AREX 2000 were drawn from the following sources:

- Internal Revenue Service (IRS) Individual Master File
- IRS Information Returns Master File
- Department of Housing and Urban Development (HUD) Tenant Rental Assistance Certification System File
- Center for Medicare and Medicaid Services Medicare Enrollment Database File
- Indian Health Services Patient Registration System File
- Selective Service System Registration File
- Social Security Administration Numident File

These administrative records were processed into a prototype census-like database of address and person records called the Statistical Administrative Records System (StARS). The person records in StARS had the same characteristics as on the Census 2000 short form: age, sex, race, and Hispanic origin. Modeling was used to impute content when necessary. The national StARS data formed the foundation of AREX 2000.

The AREX 2000 was conducted in two sites. One site consisted on two counties in Maryland: Baltimore County and Baltimore City. The second site included three counties in Colorado: Douglas, El Paso and Jefferson Counties. The test sites had approximately one million housing units and two million people in the 1990 Census. Each site included areas expected to be difficult to enumerate and areas expected to be easy to enumerate.

In addition to testing the feasibility of an ARC, another goal of AREX 2000 was to determine the optimal method for conducting such a census. The first method, known as the top down method, provides population counts at the census block level. The second method, the bottom up method, attempts to match administrative records to an independent address list and reconcile differences through field operations. The bottom up method provides both population and housing unit counts. Both methods meet the data requirements for apportionment and redistricting: block-level counts of the total population by race, Hispanic origin and age.

This paper describes the two methods tested in AREX 2000, the creation of the prototype StARS database, some additional operations done in the AREX 2000 test sites, and some preliminary results of StARS and AREX 2000. Heimovitz (2002) and Judson and Bauder (2002) give more detailed results of AREX 2000.

2. Methods for Conducting AREX 2000

AREX 2000 explored two different methods for conducting an ARC. The primary difference between the two methods was in the use of an independent address list to create a universe or frame of housing units for the test sites.

The first method, referred to as the top down method, did not use an address list as a control frame. This method is called the top down method because it begins with the national administrative records data and through various processes allocates person records to lower geographic levels, ultimately to census blocks. The method did not attempt to place people in individual housing units and as a result, did not provide a traditional census of people in households and housing units. Instead, the top down method produced block-level population counts derived only from administrative records. These counts met the minimum data requirements for apportionment and redistricting by including demographic counts by age, race and Hispanic

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

origin.

The second method, the bottom up method, went a step further and attempted to match the administrative records to addresses on an independent list of residential addresses. For AREX 2000, the independent address list was the Decennial Master Address File (DMAF), the address control frame for Census 2000. For matched addresses, the enumeration data came from administrative records. Non-matched administrative records addresses were reconciled through field operations. For non-matched addresses on the DMAF, the traditional census-taking method was used. In AREX 2000, this was simulated by extracting data from Census 2000 for these non-matched addresses. One can think of this as "nonresponse followup" of addresses where administrative records are not available for enumeration.

This method is called the bottom up method because it starts with a master address list as the foundation, much like a traditional census, and attempts to enumerate the population at those addresses, first from administrative records and then from traditional data collection methods. The bottom up method also meets the minimal requirements for apportionment and redistricting. Because housing units are also enumerated as part of the bottom up method, it provides additional data that the top down method does not, such as household sizes.

3. Creating the StARS Database

The data for AREX 2000 were initially provided by the StARS database, a national census-like file of address and person records, many of which are placed in detailed geography like census blocks. The prototype StARS database was designed mainly to meet the needs of AREX 2000. These needs included comprehensive population coverage, census short-form content, relatively simple processing, and privacy and confidentiality protections among other requirements.

The files for StARS were selected with these needs in mind. Most of the U.S. population pays taxes, thus the two IRS files had the broadest population coverage, with the other files used to fill in gaps (Huang and Kim, 2000). For example, Selective Service captures young males, and Medicare includes the older population. Because the AREX 2000 schedule required that StARS be completed well ahead of the census, the prototype StARS database was built from source files that predated Census 2000 by about 15 months.

All of the source files contain a Social Security Number (SSN) for each person record. This enabled relatively simple unduplication of person records across files. This also made it possible to ensure person records were within the population universe by verifying SSNs against the Numident file, the master list of SSNs. Person records with unverified SSNs were not included in the StARS database or in AREX 2000.

The use of the SSN created potential privacy and confidentiality issues. To reduce the chances of improper access to or use of individual micro-data, the Census Bureau established a restricted access policy for internal handling of all administrative data that requires removal of names and SSNs from output files (Clark and Gates, 1999). In StARS, we removed these personal identifiers as part of the final processing to comply with this Census Bureau policy.

The administrative source files also contain an address for each person record. Part of the StARS creation process involved geocoding these addresses to census blocks, the level required for redistricting and federal funds allocation. About 75 percent of the input addresses were geocoded to a block. For city-style addresses, those with standard house numbers and street names, the geocoding success rate was 85 percent.

The source files also contained some of the content needed for AREX 2000, although race and ethnicity presented a large challenge. The federal race and ethnicity definitions have changed over time, but the administrative records have not been updated to keep up with the changes. It is too burdensome for agencies to recontact applicants who may have filled out their original forms many years or decades ago. For example, the race on many Numident records is either White, Black or Other. A complicating factor is that race and ethnicity are almost always missing for children under age 16, whose parents were not legally required to report race or ethnicity to obtain an SSN when the children were born. Race and Hispanic origin data were thus often too coarse or too frequently missing for use in AREX 2000. To overcome the limited race data, we developed a model to impute race and Hispanic origin. Bye (1999) gives details on the race model. Sex and mortality status were missing in some cases as well, and separate models imputed those characteristics.

Some person records had content or addresses that differed across the source files. For example, a person's age might be 25 in one file and 45 in another. We resolved discrepant data using a heuristic algorithm that weighed the currency and accuracy of the various source files to choose which data to keep in StARS.

The final national StARS database contained about 257 million person records associated with 105 million addresses. In contrast, Census 2000 had about 281 million person records and 120 million addresses. The time lag and imperfect coverage of the StARS source files caused much of the difference between StARS and the census. But StARS nonetheless provided a solid foundation for AREX 2000, capturing about 90 percent of the census addresses and people using only seven source files and relatively simple processing. The creation of the StARS prototype also identified key areas for improvement to use in updated iterations of the database. Farber and Leggieri (2002) detail the building of StARS and additional results from StARS.

4. AREX 2000 Operations

Using ZIP Code and geocoding information, data for the addresses and persons in the AREX 2000 test sites were extracted from StARS. Additional clerical, field and processing operations were conducted in the AREX 2000 test sites for two reasons. First, some operations were necessary to support testing of the bottom up method. Second, additional improvements to the StARS data were made that would have been too costly to implement on a national scale. The additional AREX 2000 operations were:

- clerical geocoding
- request for physical address
- match to the DMAF
- field address verification
- rematch to the DMAF

The match to the DMAF and field address verification supported the bottom up method, while the other operations applied to both methods.

4.1 Clerical geocoding

Because one of the decennial census requirements is to produce block level counts for redistricting and funds allocation, it was important to geocode as many AREX 2000 addresses as possible. Addresses that were ultimately not geocoded to a census block were dropped from the final AREX 2000 tabulations. Geocoding of the national StARS data was done entirely through computer matching. The AREX 2000 added a clerical geocoding phase to attempt to geocode those addresses in the test sites that StARS did not.

Following the StARS computer geocoding, addresses were selected and flagged for inclusion in the AREX 2000 test sites. Two different approaches were taken depending on whether or not the address was geocoded in StARS. If geocoded, an address was flagged as within the test sites if the county and block codes were among the codes known to be within the test sites. If the address was not geocoded, the address was selected by ZIP Code as a potential test site address.

After the selection of the potential test site records, addresses that were not computer geocoded were eligible for clerical geocoding through the use of Master Address File Geocoding Office Resolution (MAFGOR). MAFGOR is an existing operational capability within the Regional Census Centers (RCCs) to provide clerical geocoding. For the AREX 2000 clerical geocoding operation, addresses eligible for clerical resolution were sent to the RCCs. The total workload was 163,148 addresses. Staff in the RCCs attempted to clerically geocode these addresses using trained geographers, reference materials and maps assembled specifically for the operation. Of the addresses eligible for clerical geocoding, 49,572 (30%) addresses were geocoded. The clerical geocoding operation added about 3 percent to the geocoding rate in Maryland and about 5 percent in Colorado.

4.2 Request for physical address

Non-city-style addresses, those with a Post Office (P.O.) Box or rural route and box number, pose a special challenge when matching and geocoding addresses. For example, the holder of a P.O. Box may actually live outside of the test site but receive mail at a P.O. box within the test site. The precise location of non-city-style addresses is often very difficult to determine. In most cases, they cannot be geocoded without a field visit. To alleviate some of these difficulties, an attempt was made to obtain a physical address, meaning a house number and street name, for non-city-style addresses via a mailed questionnaire. The form also included space for drawing a map of the residence location. For physical addresses obtained in this manner, we determined whether they were in the test sites and if so, attempted to geocode them.

The request for physical address (RFPA) questionnaire was sent to 58,151 addresses associated with 138,653 person records. Of the 138,653 people, 27,738 had no other type of address listed in administrative records source files. Only 11,683 letters were returned with usable information. Of these, 9,431 provided addresses that were geocoded, and only 8,090 gave addresses that geocoded to the test sites. Based on the low response rates and small number of addresses geocoded to the test sites, we decided not to incorporate the results of the RFPA operation into AREX 2000. Berning (2002) details the RFPA process and results.

4.3 Match to the DMAF

To support the bottom up method, administrative records addresses were matched to the DMAF in the AREX 2000 test sites. Administrative records people were then assigned to DMAF addresses based on the results of the match. In some cases, inconsistencies resulting from the match needed to be resolved. There were three potential results to the administrative records and DMAF address match:

- An administrative records address matched a DMAF address. The administrative records person information for this address was used to create an AREX 2000 household. In some cases, an ungeocoded AREX 2000 address matched to a geocoded DMAF address. This enabled the bottom up method to geocode additional addresses where earlier operations had failed.

- An administrative records address did not match a DMAF address. The information for this record was reviewed and if certain criteria were met, the address was eligible for field address verification, an AREX 2000 operation described below.
- A DMAF address did not match an administrative records address. These addresses were assumed to be valid and enumeration data for the household at the address was extracted from Census 2000. This step simulated a nonresponse followup for housing units not enumerated by administrative records.

The DMAF match had a computer phase followed by a clerical phase. The clerical phase included a review of possible computer matches and an attempt to match addresses that were not matched in the computer phase. About 80 percent of the eligible AREX 2000 addresses were matched to the DMAF during the computer phase. The clerical phase added an additional 4 percent, making the final match rate about 84 percent in the AREX 2000 test sites.

4.4. Field address verification

We implemented the field address verification (FAV) operation to check the validity of administrative records addresses that were not matched to the DMAF following the computer and clerical matching operation.

To minimize the amount of field work required, the assumption was made that any non-matched DMAF addresses were, in fact, valid and existent because of the numerous operations that went into building the DMAF. As a result, only non-matched administrative records addresses were eligible for FAV. Due to resource constraints, a sample of addresses was selected across the test sites for field verification.

The universe of addresses eligible for FAV was restricted to geocoded, city-style addresses within the AREX 2000 test sites. The universe excluded some ZIP Codes that belonged to three colleges, a medical center, and an Air Force base in the assumption that these were group quarters that included few or no residential addresses. The number of addresses to be verified was based on a stratified cluster sample of unmatched, city style addresses. The sample resulted in the selection of 6,644 addresses for the FAV operation.

Because field staff were already committed to decennial operations, volunteers from headquarters were recruited to conduct the field verification. The twenty volunteers were divided into two teams, one for each site. To prepare for the field operation, a two-day training seminar was conducted. In addition to classroom training, teams were given an assignment for a residential area near the Census Bureau. Results of the field training were reviewed and volunteers debriefed prior to certification. A set of maps was also produced for each area. For each address selected as part of the sample, an address listing page was printed. The listing page included a series of yes or no questions to determine if the address actually existed and to collect intelligence about the address if it was found. About 30 percent of the sampled addresses were found to exist in the field. These results were applied within strata to estimate the overall percentage of addresses in the FAV sample universe that actually existed.

4.5 Rematch to the DMAF

A final match of the AREX 2000 addresses to the DMAF was made to transform collection geography to tabulation geography. The census is conducted using collection geography, codes assigned to areas by the Census Bureau. Census results are presented in tabulation geography, defined to reflect political and statistical boundaries. Because the AREX 2000 addresses were initially geocoded to collection geography, it was necessary to translate to tabulation codes to enable comparisons to Census 2000.

During the rematch, a problem with multiple DMAF identifiers surfaced. The dynamic nature of the DMAF required that DMAF identifiers be continually updated from census operations. Thus, the number of multiple identifiers for a given address may have changed since the first computer match.

The impact of this on the correctness of tabulation block assignments immediately prior to creating the AREX 2000 results is difficult to assess. We may conduct further research into multiple DMAF identifiers and their effects on the AREX 2000 results.

5. Preliminary Results

The preliminary results of the two AREX 2000 methods are shown in Tables 1 and 2 below.

Table 1. Bottom Up Population Totals

Test Site County	AREX 2000	Census 2000	$\frac{AREX}{Census}\times 100\%$
Baltimore City, MD	661,561	651,154	102%
Baltimore County, MD	745,893	754,292	99%
Douglas County, CO	170,102	175,766	97%
El Paso County, CO	509,597	516,929	99%
Jefferson County, CO	508,254	527,056	96%

Table 2. Top Down Population Totals

Test Site County	AREX 2000	Census 2000	$\frac{AREX}{Census}\times 100\%$
Baltimore City, MD	570,648	651,154	88%
Baltimore County, MD	696,183	754,292	92%
Douglas County, CO	148,270	175,766	84%
El Paso County, CO	456,891	516,929	88%
Jefferson County, CO	473,495	527,056	90%

These preliminary results strongly indicate that the combination of an ARC and a traditional census in the bottom up method yields more accurate results. More detailed results are given in Heimovitz (2002) and Judson and Bauder (2002).

Note that AREX 2000 is part of the Census 2000 Testing and Experimentation program and thus requires a full review before final results are presented. The results presented here should be regarded as preliminary. Also, the results in Tables 1 and 2 exclude group quarters, which causes these results to differ from some results in other AREX 2000 papers.

6. Conclusions

One of the experiments implemented for Census 2000 was AREX 2000. This was a first attempt to simulate an ARC in the United States. The StARS prototype, a nationwide database of both person and address information from seven federal administrative records files, was built to facilitate the experiment. StARS was created by merging and unduplicating the input files, geocoding the address information to census blocks and imputing missing data items. AREX 2000 was an operational success. File acquisition, clerical, field and processing operations were completed to yield a simulation of a census based on administrative records.

This nascent exploration of administrative records for census taking reveals needed improvements in coverage and content as well as refinements in processing and operational components. While the Census Bureau does not expect to replace traditional census taking methods with administrative data, this experimental research opens the door to opportunities for supplementing the census through evaluation, targeting, and imputation as well as other non-decennial census applications such as improvements to demographic surveys, intercensal population estimates and the Master Address File.

7. References

Berning, M.A. (2002), “Administrative Records Experiment in 2000 (AREX 2000): Request for Physical Address Evaluation,” Internal Census Bureau memorandum.

Bye, B. (1999), “Race and Ethnicity Modeling with SSA Numident Data,” Internal Census Bureau memorandum.

Clark C. and Gates, G. (1999), “Memorandum on Restricted Access Policy for Administrative Records,” Internal Census Bureau memorandum.

Farber, J.E. and Leggieri, C.A. (2002), “Building and Validating a National Administrative Records Database for the United States,” Administrative Records Memorandum Series, Washington, DC: U.S. Census Bureau.

Heimovitz, H.K. (2002), “Administrative Records Experiment in 2000: Outcomes,” paper presented at the Joint Statistical Meetings.

Huang, E. and Kim, J. (2000), “One percent Sample Study Report,” Administrative Records Research Memorandum Series #42, Washington, DC: U.S. Census Bureau.

Judson, D.H. and Bauder, D.M. (2002), “The Administrative Records Experiment in 2000: Evaluating the Ability of Administrative Records Databases to Replicate Census 2000 Results at the Matched Household Level,” paper presented at the Joint Statistical Meetings.