Working Paper No. 5
ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE-EUROSTAT Work Session on Registers and
Administrative Records for Social and Demographic Statistics
(Geneva, 9-11 December 2002)

Invited paper

MERGING ADMINISTRATIVE RECORDS DATABASES IN THE ABSENCE OF A REGISTER: DATA QUALITY CONCERNS AND OUTCOMES OF AN EXPERIMENT IN ADMINISTRATIVE RECORDS USE

Submitted by Dean H. Judson[1], U.S. Census Bureau

[1] 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. In addition, this paper represents the views of the author and does not necessarily represent the views of the US Census Bureau or any other agency of the US government. Data presented here are subject to change in the Quality Assurance process.

I. INTRODUCTION: POTENTIAL USES OF ADMINISTRATIVE RECORDS AT THE U.S. CENSUS BUREAU

1. 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. In the U.S., interest in taking a decennial census by administrative records dates back at least as far as a proposal by Alvey and Scheuren (1982) that records from the Internal Revenue Service (IRS) along with those of several other agencies might form the core of an administrative record census. The National Research Council recommended that research proceed on both the "small area estimates" and "administrative records census" fronts.

2. On the first front, recent proposals from the Census Bureau itself, including the 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 (e.g. Citro and Kalton, 2000a,b; Fisher and Campbell, 2002; Popoff, O'Hare, and Judson, 2002).

3. At the same time, on 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).

4. The purpose of this paper is to document administrative records use for decennial purposes at the Census Bureau both historically and currently, and identify successes and current challenges. (A much broader discussion, including areas outside the decennial census, is provided in Judson, 2001. More detailed descriptions of these results can be found in Leggieri, Pistiner, and Farber, 2002; Heimovitz, 2002a,b; Judson and Bauder, 2002a,b; and Cook and Berning, 2002. A long-range research agenda can be found in Leggieri and Prevost (1999)).

II. WHAT ARE ADMINISTRATIVE RECORDS?

5. According to the American Statistical Association Ad Hoc Committee on Privacy and Confidentiality (1977), "Administrative 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.

III. ADMINISTRATIVE RECORDS CENSUS EXPERIMENT IN 2000 (AREX 2000)

6. 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 development 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.

7. 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.

8. The Administrative Records Experiment 2000 (AREX 2000) is a simulation of an Administrative Records Census (ARC). It was being conducted in conjunction 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 (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.

9. In order to perform an ARC simulation, the following steps had to take place: The sites for the simulation had to be chosen; the administrative databases that would have the highest overall population coverage needed to be defined and obtained; these files needed to be edited to standardize their concepts and, where necessary, translate into census concepts; field operations needed to be defined and executed; and, of course, post-processing of the composite files needed to be performed. AREX 2000 accomplished these operations.

IV. AREX TOP-DOWN AND BOTTOM-UP METHODS

10. Knott (1994) identified two basic ARC models: (1) the top-down model that assembles administrative records from a number of sources, unduplicates them, assigns geographic codes and counts the results; and (2) the bottom-up model that links administrative records to a master address file, fills the addresses with individuals, resolves gaps and inconsistencies address by address, and counts the results. Knott also suggested a composite top-down/bottom-up model. It would unduplicate administrative records using the SSN then link the address file and proceed as in the bottom-up approach. In overall concept, AREX 2000 most closely resembles this composite approach.

11. The AREX 2000 enumeration can be conceived as occurring in two phases. The first, or Top-down, phase involved the assembly of records from a number of national administrative record systems and unduplication of individuals within the combined systems. This was followed by computer geocoding of street addresses to the level of Census block and two attempts to obtain and code physical addresses for those that would not geocode by computer. Finally, there was a selection of “best” demographic characteristics for each individual and “best” street address within the experimental sites.

12. One can think about the results of the Top-down process in two ways. First, counting the population at this point provided, in effect, an administrative-records-only census. That is, the enumeration included only those individuals found in the administrative records, and there was no other support for the census outside of activities related to address coding. However, without a national population register as its base, one might expect an enumeration that used only administrative records to be substantially incomplete. And so a second way to think about the top-down process is as a substitute (or analogue) for an initial mail-out in the context of a more conventional census that would include additional support for the enumeration.

13. The second phase of the AREX 2000 design was an attempt to complete the administrative-records-only enumeration by the correction of errors in administrative records addresses through address verification (a coverage improvement analogue) and by adding persons missed in the administrative records (a non-response follow-up analogue) . This phase began by (computer and clerically) linking the addresses found in the top-down process to the Master Address File (MAF) in order to assess their validity and to

identify those MAF addresses for which no administrative records were found. There were effectively three outcomes of the linkage: 1) Administrative Record (AR) address links to a MAF address; 2) AR address is not linked to a MAF address; 3) MAF address remains unlinked to any AR address.

- For AR addresses linked to a MAF address, the administrative records enumeration was used without adjustment.
- For AR addresses not linked to a MAF address, A field address verification (FAV) was used to verify these addresses indeed exist, and invalid administrative records addresses were excluded from the bottom-up best address selection process.
- For MAF addresses not linked to the AR list, these MAF addresses were canvassed in order to enumerate persons at addresses not found in the administrative records systems. In the AREX, this canvassing was simulated by adding those persons found in the 2000 Census at the unlinked addresses to the adjusted administrative-records-only counts, thus “completing” the enumeration. Doing the AREX as part of the 2000 Census obviated the need to mount a separate field operation to canvass the unlinked MAF addresses.

14. Considering the Top-down and Bottom-up processes as part of one overall design, AREX can be thought of as a prototype for a more or less conventional census with the initial mailout replaced by a top-down administrative records enumeration, or an administrative records census with conventional-style enhancements.

V. LIMITATIONS

15. There were four principal limitations on the experiment.

- The administrative records source files were limited to those used in the creation of StARS 1999 which relied primarily on files for tax year 1998 and other files extracted early in calendar year 1999. These files neither exhausted the national-level administrative records that might have been available for AREX 2000 nor were they the most timely with respect to April 1, 2000, Census Day for the 2000 Census.

- The number of experimental sites was small. Although it would not have been reasonable or realistic to attempt to mount this first AREX in a representative sample of geographic areas large enough to make national estimates, additional sites would have provided more confidence that the results were not idiosyncratic to the sites selected.

- There was no experimental variation in key design parameters such as the clerical and field operations and the address selection algorithm. Without some factorial or fractional factorial structure, direct estimates of operational impacts of components, individually or in combination, were not possible.

- The measurement of race and Hispanic origin in administrative records at the national level is deficient. Attempts were made to improve the measurement through the use of certain statistical models (Bye, 1998; Bye and Thompson, 1999), but the results were not entirely satisfactory.

16. The limitations in the AREX were largely due to resource constraints and a short planning period for what was an extremely complex and novel undertaking. In evaluating the AREX, it is important to understand that processing difficulties and weaknesses in the outcomes are not insurmountable.

VI. EXPERIMENTAL SITES AND DATA SOURCES

17. Two sites were selected that had a total of approximately one million housing units and a population of approximately two million persons. One site included Baltimore City and Baltimore County, Maryland. The other site included Douglas, El Paso and Jefferson Counties, Colorado. The sites were

purposively chosen to provide a mix of population and housing characteristics needed to assess the difficulties that might arise in conducting an administrative records census.

18. StARS 1999, the national administrative records data base developed by ARRS, was the source of the administrative records for the experiment (see Judson, 2000, for details). The national level files that contributed to the StARS 1999 database and therefore to AREX 2000, were:

- Internal Revenue Service (IRS) Tax Year 1998 Individual Master File (IMF 1040),
- IRS Tax Year 1998 Information Returns File (IRMF W-2 / 1099),
- Department of Housing and Urban Development (HUD) 1999 Tenant Rental Assistance Certification System (TRACS) File,
- Center for Medicare and Medicaid Services (CMS) 1999 Medicare Enrollment Database (MEDB) File,
- Indian Health Services (IHS) 1999 Patient Registration System File, and
- Selective Service System (SSS) 1999 Registration File.

19. The cutoff date for the IRS 1040 file was 9/30/99. For all other files the cutoff date was 4/1/99. The gap between file extraction and the 2000 Census Day had an impact on both population coverage--births, deaths, immigration and emigration--and geographic location--housing extant, and geographic mobility.

20. The source of most of the demographic data and some of the death data was the Census Numident, an edited version of the Social Security Administration's (SSA) Numerical Identification (Numident) File. The SSA Numident file is the numerically ordered master file of assigned Social Security Numbers (SSN) that may contain up to 300 entries for each SSN record, although on average contains two records per SSN. Each entry represents an initial application for a SSN or an addition or change to the information pertaining to a given SSN. The SSA Numident available for StARS 1999 reflected all transactions through December 1998. The Census Numident was designed to collapse the SSA Numident entries to reflect "one best record" for each SSN containing the "best" demographic data--date of birth, sex, race, Hispanic origin--for each SSN on the file. Missing or unknown race and Hispanic origin was imputed using statistical models.

VII. TOP-DOWN ENUMERATION

21. To use administrative records to identify individuals residing at geocoded addresses in the AREX test sites and to assemble the necessary demographic data, a "dual-stream" processing approach was adopted. One processing stream produced a unique record for each individual with best demographics. The second stream produced a set of addresses, geocoded to the block level. In the end, persons and addresses were brought together, a best address was selected for each person to complete the Top-down enumeration. Much of the data processing to accomplish these tasks was accomplished in the development of StARS 1999.

22. Person processing consisted of three main steps: (1) file edits for person data, (2) SSN verification of person records, and (3) unduplication of person records using the SSN, and selection of the "best" demographic characteristics for each person record. Top-down Address Processing consisted of four main steps: (1) file edits for address data, (2) standardizing and computer geocoding the address records, (3) geocoding addresses not coded by computer, and (4) creation of a Master Housing File for administrative record addresses.

23. The computer geocoding operation was designed to identify addresses in the AREX sites and obtain Census block codes from the Topologically Integrated Geographic Encoding and Referencing (TIGER) database program. About 83 percent of the 1.3 million AREX addresses were successfully assigned a block code by TIGER.

24.   Two attempts were made to manually geocode Census block for addresses in the test sites that were not coded by TIGER. One process involved manual coding of city style addresses via MAFGOR, an existing operational capability within the Regional Census Centers. This added an additional 3 percent to the coded addresses. The other process involved obtaining physical address information for non-city style and Post Office Box addresses by a mailout/mailback procedure and then manually geocoding the addresses. The mailing was sent to 58,151 addresses associated with 138,653 individuals. For a number of reasons, the response rate to the mailout was only about 20 percent of which about 8,090 were geocoded to an AREX test site county. Consequently, these results were not used in the final AREX results.

25.   The creation of a Master Housing File of unduplicated administrative record addresses was the final step in the address processing before the addresses were relinked with the person records. At this point address and person data were brought together in preparation for creation of a composite person record. There were two principal tasks. First, individuals potentially in the AREX test sites were identified. Then, the best address was selected for these persons. If the best address was in the test site, then the individual became part of the Top-down enumeration.

VIII. BOTTOM-UP ENUMERATION

26.   The Bottom-up phase of the AREX 2000 design was an attempt to complete the administrative-records-only enumeration by adding persons missed in the administrative records, a process analogous to a conventional non-response follow-up (NRFU). There was also an attempt to correct Top-down enumeration errors by removal of invalid administrative records addresses prior to best address selection. A valid address was defined as one that was linked to a MAF address or was unlinked but deemed valid after a field address review. There was no provision for correcting enumerations at households with valid addresses.

27.   The Bottom-up operational components of AREX were conducted on records contained within the five test site counties. These operations consisted of:

i)   Computerized record linkage of the AREX addresses to the MAF.
ii)   Clerical review and attempted reconciliation of unlinked AREX addresses.
iii)   Field Address Verification (FAV) of unreconciled administrative record addresses and address re-selection.
iv)   Census Pull, the simulated NRFU for unlinked Census 2000 addresses.
v)   Bottom-up enumeration.

28.   To most accurately link administrative records addresses to the DMAF, the link was limited to AREX addresses which were geocoded or with a standardized street name, a standardized property description or both. Excluded from the linking process were non-standardized addresses, standardized post office or box addresses, standardized post office and rural route addresses and undefined addresses. The file to which the AREX addresses were linked consisted of a list of addresses on the DMAF whose current county code showed the address to be within one of the five AREX test site counties.

29.   The linking process consisted of several passes of AutoMatch, a commercial software package that applies probabilistic record linkage techniques. About 80 percent of the 1.2 million eligible administrative records addresses linked the DMAF. A clerical review of the non-linked addresses added an additional 4 percent to the computer link process.

30.   The Field Address Verification operation was implemented to check the validity of administrative records addresses that remained unlinked to the DMAF following the computer linking and clerical review, sometimes by correcting erroneous address field values. The original plan called for a review of 100 percent of the unlinked addresses by Census field staff, but the plan was changed to have only a sample of addresses reviewed by central office volunteers. The results from the sample were used to estimate a regression equation giving the probability of a valid address. The equation was then used to impute validity or lack thereof to the non-sample addresses.

31. A sample of about 6600 addresses was selected from the 153,000 addresses eligible for FAV. About 30 percent of the sample addresses were found to be valid, one quarter valid as listed and the remainder valid after corrections. The net effect of the FAV sample and imputation was that about 93,000 administrative records addresses were removed from consideration prior to Bottom-up address reselection.

32. For the Bottom-up enumeration, the NRFU analogue was simulated by including persons found in the Census 2000 Hundred Percent Detail File (HDF) at addresses that were not among the final set of "best" administrative records addresses. These were persons enumerated in Census 2000, and the assumption was that they would have been counted in the AREX had some sort of follow-up been instituted. The process of including persons from the Census 2000 HDF was referred to as the Census Pull and resulted in the inclusion of about 312,000 persons at 117,000 non-vacant HDF addresses. There was no attempt to unduplicate persons after the Census Pull.

IX. OVERALL AREX ENUMERATION RESULTS

33. The AREX enumeration results are shown in Table 1 (see also Berning and Cook, 2002, and Leggieri, Pistiner, and Farber, 2002). As expected, the Bottom-up coverage is much improved compared to the Top-down, and this is largely due to the completion of the Top-down enumeration by the Census Pull, simulating a follow-up to the administrative records enumeration. Specifically, the Bottom-up coverage of children (81% - 94% across the test sites) is substantially better than the Top-down (72% - 83%). Coverage of children is a particular weakness for administrative records used in AREX 2000.

34. Adults in the Bottom-up are more or less uniformly overcounted (102% - 104%), and figure one illustrates that the overcounts increase with age. The overcount of adults most likely is due to unaccounted for deaths in the 12 months prior to Census Day, the lack of special populations operations in the AREX, and failure to unduplicate persons after the Census Pull. Of course, the latter means that there is some duplication of children as well.

35. Two of the Bottom-up operations entailed an attempt to improve the administrative records addresses prior to Bottom-up "best" address selection: (1) the initial link to the DMAF and its follow-up clerical review, and (2) the FAV. The following tables provide some information on the net impact of these operations on Bottom-up person tallies from administrative records, and in doing so, provides a comparison of the Top-down and Bottom-up results with regard to administrative records processing.

36. Of the 2.3 million persons tallied in the Top-down enumeration, 70,031 (about 3%) were excluded from the Bottom-up administrative record counts. These exclusions occurred either because the only address that the persons had was rejected by the Bottom-up processes or because the only remaining addresses were outside of the AREX test sites. It is possible that many of the 70,000 persons that were excluded from the Bottom-up administrative record tallies could have been reintroduced to the Bottom-up enumeration by the Census Pull, if they in fact were resident in the AREX test sites.

37. Table 2 shows 7,525 administrative records persons included in the Bottom-up that were not in the Top-down. Evidently, these cases were acquired as a result of block coding for some administrative records addresses in the Bottom-up processes that were not coded prior to the initial DMAF link or due to FAV sample address changes.

38. As Table 3 illustrates, for administrative records persons enumerated in both the Top-down and the Bottom-up, the additional operations seem to have had very little impact; over 99% of these persons were at the same address in both enumerations. Overall, the net effect of the Bottom-up operations on the administrative record tallies was quite modest in that more than 96% of the persons enumerated in the Top-down process were enumerated at the same address in the Bottom-up process.

X. DETAILED AREX ENUMERATION RESULTS

39. Detailed enumeration results focused mainly on a comparison of the Bottom-up enumeration with the Census 2000 (Heimovitz, 2002a,b). The analysis did not include group quarters or Census 2000 respondents with “multi” or “other” race. The analysis progressed from large geographic areas to small geographic areas, beginning with the five test site counties and ending with Census blocks within the sites. The evaluation incorporated a variety of methods to accomplish its objectives, including univariate and multivariate statistical analyses of AREX-Census differences, and spatial/ecological maps that examined the geographic distributions of key comparison measures. The outcomes evaluation tried to disentangle the influence of demographic change and AREX processing, coverage and data quality issues, while presenting basic enumeration statistics.

40. At the county level, the Bottom-up process undercounted total population in all sites except Baltimore City. As with the total population, males and females were undercounted in all sites except Baltimore City, but the female undercounts were slightly greater than male undercounts. Age groups showed more variability with most groups undercounted. Generally the size of the undercounts increased with decreasing age, except for the 20-24 age group (see figure one). These patterns did not appear to be site-specific. Overcounts for the oldest old and undercounts for the youngest persons suggest that much more timely birth and death information must be obtained, especially from the Social Security Administration and the Center for Medicare and Medicaid Services. Also, the special enumeration needs of populations such as college students, the military and persons in nursing homes must be incorporated into administrative records processes.

41. Administrative records were not a good source for race and Hispanic origin, and the models were not strong enough to cover their deficiencies. Blacks and Hispanics were undercounted when they were a large minority group and overcounted when they were not (see figure two). American Indians and Alaskan natives were not well identified and the accuracy of Asian/Pacific Islander counts was uncertain. Bottom-up state legislative districts compared remarkably well with Census 2000. For both sites, a greater number of districts were overcounted rather than undercounted because of Bottom-up overcounts of voting age adults. However, differences for both overcounted and undercounted districts were small in magnitude.

42. Bottom-up tract-level total population results indicated a reasonable correspondence between AREX and Census. The population counts of 70% of tracts were within 5% of the Census total, and 95% of the tracts were within 25% of the Census total, though a sizable number of tracts had moderate and large undercounts. At the block-level, population counts were the least accurate. For the total population only 38% of blocks met the 5% criterion and about 85% of blocks met the 25% criterion.

43. A multivariate analysis of block differences showed that large undercounts were associated with such block characteristics as high population density, high rental rates, and large proportions of persons age 20-24 or 65 and over. Large overcounts were associated with high vacancy rates, low population density, small proportions of persons under the age of 19 and large proportions of persons age 20-24 and 65 and over.

XI. HOUSEHOLD-LEVEL ANALYSIS

44. The general goal of the household-level analysis (Judson and Bauder, 2002a,b) was to assess how well households formed from administrative records “demographically matched” those from Census 2000 at the same addresses. The evaluation focused, first, on the factors associated with linked AREX and Census 2000 HDF addresses. Then, demographic comparisons were made between households at linked addresses. There was a special focus on Census 2000 households that required a non-response follow-up and Census 2000 unclassified (imputed) households.

45. The evaluation used both descriptive analyses and logistic regression analysis to assess the coverage and accuracy of AREX households. Descriptive analyses were performed for households in all five AREX counties and for the Census 2000 NRFU and imputed households in the test sites. A regression

model was developed to predict the probability of an accurate household demographic match using address and AREX processing characteristics as predictors. Addresses with a high probability of correct demographic match between occupants might be candidates for administrative records substitution in the case of NRFU or imputation in a conventional census. In the following discussion the term “linked” is used to mean a computer/clerically linked address. The term “matched” or “demographically matched” is reserved for household demographic comparisons at linked addresses. A linked pair of households “demographically match” when they have the same number of persons and their age, race, sex and Hispanic origin distributions are equivalent.

46. AREX’s coverage of the Census NRFU universe was not as good as its coverage of the non-NRFU universe. Table 4 illustrates that AREX housing units were linked with 70.9 percent of the Census NRFU housing units, compared with 88.4 percent of the Census non-NRFU housing units. For occupied NRFU housing units, the coverage rate was 76.7 percent. AREX housing units were linked with 63.2 percent of households that were imputed to have people in them, and 34.7 percent of those imputed to be vacant.

47. Again table 4 indicates that AREX and Census counted the same number of people in the housing unit for 51.1 percent of the 889,638 linked households, and AREX was within one of the Census for 79.4 percent of the units. The 51.1 percent is effectively a ceiling on the percent of linked households that had exactly the same persons from AREX and Census 2000. Although errors in address linkage would account for some of the mismatched households, the deficiencies in administrative records cited earlier in this report--missing children, lack of special population operations and the time gap between the administrative records extracts and Census day--most likely account for the major part.

48. Again in table 4, for linked NRFU housing units, AREX had the same numbers of persons for 37.0 percent of the units and was within one only 69.3 percent of the time. Evidently, Census 2000 NRFU housing units were more susceptible to AREX deficiencies than non-NRFU units. In addition, enumeration errors in Census 2000 might have been higher for these units. A similar relationship holds between Census imputed addresses and AREX, although by definition because these addresses were imputed we naturally expect demographic details to vary. However to the extent that AREX addresses contain the same number of persons, this supports the veracity of Census imputation methods. Finally, table 5 illustrates a basic relationship between AREX and Census households: As the census household size increases, the percentage of AREX addresses that “demographically match” decline.

49. As table 6 illustrates, having all persons in the household aged 65 or older is a very strong predictor of demographic matching (71.57 percent), as opposed to other households (33.44 percent). We do not know if this is because of better data quality for persons 65 or older, the Medicare source file for many of these persons, lower mobility rates of such addresses, or more accurate census responses by such persons. These are all conceivable explanations for this effect.

50. Logistic regression analysis revealed a number of factors associated with greater probability of matched household demographics. These include: single unit address rather than multi-unit, household with only 1 or 2 members, all household occupants over the age of 65, at least one White occupant, no occupant with imputed race in the AREX. The predictive power of the model was moderately strong. At a predicted probability of 0.5 or higher, the probability of a correct classification was about 72 percent. Evidently, the limitations in the data, particularly the administrative records cutoffs and poor race and Hispanic origin measurement, made household prediction difficult.

XII. IMPLICATIONS FOR 2010 PLANNING

XII.1 There should be an administrative records census experiment in 2010

51. AREX 2000 was certainly too small and limited to conclude that a decennial census based largely on administrative records would be viable in 2010. Nevertheless, the success of the AREX despite its limitations strongly suggests that research on the possibility of an administrative records census continue apace, partly for its own sake, but even more so for the variety of useful applications of administrative

records research for many Census Bureau operations. Therefore, if the U.S. Census Bureau wishes to continue to evaluate the poential for an ARC, there should be an ARC Experiment in 2010 (AREX2010) that would address the key deficiencies of AREX 2000.

52. The experiment should be of sufficient size and scope so that generalizations to the entire 2010 population are possible. The kind of information needed to be sure that ARC methodology would produce an accurate enumeration, and to convince others of this, cannot be obtained from a small, geographically limited experiment.

53. The administrative records used in AREX 2010 would have to be more timely, relative to Census Day, than those used in AREX 2000. This appears to be technically feasible if certain administrative records extracts are more current and other administrative data are received on a flow basis in the year of the census. Also, there could be additional sources not used in AREX 2000 that would fill some of the coverage gaps in the AREX 2000 record set.

54. The enumeration of special populations in an ARC context must be fully implemented in the 2010 test. This includes not only the well known areas such as the military, college students and various group quarters, but might also include other populations such as children who are not covered well in national level administrative records.

55. The measurement of race and Hispanic origin must be improved since they are crucial to the success of a short form census. Fortunately there is a methodology currently being explored at the Census Bureau that should provide the much needed improvement.

XII.2 Planning and research in support of arex 2010 should begin now

56. One thing learned from the AREX 2000 experience is that a planning period of 2-3 years is not sufficient to mount a full ARC test. Negotiations alone for additional and more timely administrative data are likely to take several years. While the use of administrative data is an important part of the Bureau's mission, there is often little benefit to the source agencies for providing such data; and absent a legal mandate to support Census activities, negotiations can be quite difficult. In addition, much research needs to be done in both person and address processing prior to the specification of an AREX 2010 design.

XII.3 Substitution for 2010 NRFU households should continue to be explored

57. Although the results of the household-level analysis were not definitive due to the limitations on AREX 2000, they were sufficiently strong that research into the substitution of administrative records households for NRFU or unclassified households in a conventional census should continue. For NRFU households there is the potential for significant cost savings, and for unclassified households, the potential for improved accuracy in addition to that provided by current hot-deck imputation.

58. The approach piloted in AREX 2000 could be tested as part of the 2010 Census Test in 2004 using models developed from a linkage of StARS 2000 data to Census 2000 files. The timing of the administrative records in StARS 2000 would be much closer to the 2000 Census Day than the StARS 1999 data used in AREX 2000, and much more like the data that could be acquired for 2010.

XII.4 Other 2010 impacts should be considered

59. There are other aspects of 2010 census development in which administrative records might play a supporting role. These include Master Address File improvements (Farber and Shaw, 2002), development and testing of unduplication methods for 2010 (Bean and Bauder, 2002), and triple system estimation research (Biemer, Brown, Judson, and Wiesen,2001; Asher and Fienberg, 2002; Stuart and Zaslavsky, 2002).

XII.5 2010 data acquisition and research agenda

60. Arrangements need to be made to acquire administrative records on a more timely basis and to obtain some datasets that might fill some of the administrative records coverage gaps. Getting more timely data will mean arranging for the receipt of IRS 1040 and 1099 records on a flow basis and possibly obtaining W-2 data from SSA rather than IRS. It will also mean obtaining more timely extracts from other agencies including SSA, CMS, HUD and IHS. New data to fill coverage gaps could come from SSA's benefit payment systems and enumeration at birth files and CMS's Medicaid eligibility files may contain missing children.

61. A research agenda for 2010 should include:

Additional evaluation of the impact of clerical and field operations in AREX 2000;
Person unduplication in the AREX Bottom-up process;
Repeating AREX 2000 analyses with StARS 2000 (instead of 1999) data;
Repeating the Household-level analysis using StARS 2000 (vs. 1999) data;
Analysis of administrative records coverage gaps;
Master Address File improvements using administrative records;
Improving address linkage techniques;
Enhancing Numident race and Hispanic origin data using Census 2000.

XII.6 Benefits for other Census Bureau programs should continue to be explored.

62. The research that went into the development of StARs and AREX 2000 has had significant payoffs in Census programs other than the decennial census, and the development of new uses for administrative records should continue to benefit non-decennial programs in the future. There have been huge gains in knowledge of the strengths and weaknesses of national administrative records systems to support various Census Bureau activities, in the capacity for large scale data processing, data standardization, record linkage (e.g., between the Current Population Survey and Decennial files), file unduplicaton, SSN search and verification, small area synthetic estimation methods, and several other applications, that will have benefits throughout the Census Bureau.

ACKNOWLEDGEMENTS

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

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

REFERENCES

Alvey, Wendy and Scheuren, Fritz (1982). Background for an Administrative Record Census. Proceedings of the Social Statistics Section, Washington DC: American Statistical Association, 1982.

Asher, Jana, and Fienberg, Stephen (2002). The Administrative Records Experiment In 2000: An Application To Population Count Estimation Via Triple Systems Estimation. To appear in the 2002 Proceedings of the American Statistical Association, Government Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

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

Bean, Susanne, and Bauder, Mark (2002). Census and Administrative Records Duplication Study. Unpublished document available from the U.S. Census Bureau.

Biemer, Paul, Brown, G. Gordon, Judson, D.H., and Wiesen, Christopher (2001). Triple System Estimation with Erroneous Enumerations in the Administrative Records List. Paper presented at the 2001 Joint Statistical Meetings, Washington DC. Under review at the Journal of Official Statistics.

Berning, Michael A., Cook, Ralph H. (2002). Administrative Records Experiment in 2000 (AREX 2000): Process Evaluation. Planning, Research and Evaluation Division Census 2000 Experiment Evaluation (DRAFT), Bureau of the Census, November, 12, 2001.

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.

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.

Citro, Constance, and Kalton, Graham (2000a). Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond. Washington, DC: National Research Council.

Citro, Constance, and Kalton, Graham (2000b). Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology. Washington, DC: National Research Council.

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, Schirm, Allen L. (1997). “On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population.” Washington, DC: Internal Revenue Service.

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

Farber, Jim, and Shaw, Kevin M. (2002). Dual System Estimates of Housing Units Based on Administrative Records. To appear in the 2002 Proceedings of the American Statistical Association, Government Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

Fisher, Robin and Campbell, Jennifer (2002). Health Insurance Estimates for States. To appear in the 2002 Proceedings of the American Statistical Association, Government Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

Heimovitz, Harley K (2002a). Administrative Records Experiment 2000: Outcomes. To appear in the 2002 Proceedings of the American Statistical Association, Government Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

Heimovitz, Harley K (2002b). Administrative Records Experiment 2000: Outcomes. Planning, Research and Evaluation Division Census 2000 Experiment Evaluation (DRAFT), Bureau of the Census, November, 12, 2001.

Judson, D.H. and Popoff, Carole L. (1998). Research Use of Administrative Records. Monograph manuscript in process available from the authors.

Judson, Dean H. (2000).The Statistical Administrative Records System: System Design, Successes, and Challenges. Presented at the NISS/Telcordia Data Quality Conference, November 30-December 1, 2000.

Judson, D.H., Popoff, Carole L., and Batutis, Michael (2001). An Evaluation of the Accuracy of U.S. Census Bureau County Population Estimation Methods. Statistics in Transition, 5:185-215.

Judson, D.H., and Bauder, Mark (2002a). Evaluating the Ability of Administrative Records Databases to Replicate Census 2000 Results at the Household Level. To appear in the 2002 Proceedings of the American Statistical Association, Government Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

Judson, Dean H. and Bauder, Mark (2002b). Administrative Records Experiment in 2000 (AREX 2000): Household Level Analysis. Planning, Research and Evaluation Division Census 2000 Experiment Evaluation (DRAFT), Bureau of the Census, November, 12, 2001.

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.

Leggieri, Charlene, and Prevost, Ron (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.

Leggieri, Charlene, Pistiner, Arona, and Farber, Jim (2002). Methods for Conducting an Administrative Records Census Experiment in 2000. To appear in the 2002 Proceedings of the American Statistical Association, Government Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

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].

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.

Popoff, Carole, O’Hara, Brett and Judson, D. H. (2002). Estimating the Proportion of Uninsured Persons at the County Level: Exploring the Use of Additional Covariates in a Synthetic Estimates System. To appear in the 2002 Proceedings of the American Statistical Association, Government Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

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.

Stuart, E. and Zaslavsky, A.M. (2002). Using administrative records to predict census day residency. In Constantine Gatsonis, Robert E. Kass, Alicia Carriquiry, Andrew Gelman, David Higdon, Donna K. Pauler, Isabella Verdinelli (Eds.), Case Studies in Bayesian Statistics Volume VI. New York, NY: Springer.

Zanutto, Elaine (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, DC: U.S. Census Bureau.

Zanutto, Elaine, and Zaslavsky, Alan M. (1996). Modeling Census Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Followup, to Impute Census Non-respondents. 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. In R. Groves, R.J.A. Little, and J.Eltinge (Eds), Survey Nonresponse. New York: John Wiley.

Annex Tables

Table 1. Basic AREX Enumeration Results

Test site and County	Census Population	Top-down Population	% of Census Population	Bottom-up Population	% of Census Population
Baltimore City Maryland	651,154	570,648	88%	661,561	102%
Under 18	161,353	134,471	83%	151,411	94%
18 and over	489,801	436,127	89%	510,109	104%
Baltimore County Maryland	754,292	696,183	92%	745,893	99%
Under 18	178,363	146,012	82%	154,500	87%
18 and over	575,929	550,086	96%	591,313	103%
Douglas County Colorado	175,766	148,270	84%	170,102	97%
Under 18	55,477	40,085	72%	46,394	84%
18 and over	120,289	108,165	90%	123,689	103%
El Paso County Colorado	516,929	456,891	88%	509,597	99%
Under 18	142,480	110,504	78%	121,647	85%
18 and over	374,449	346,322	92%	387,888	104%
Jefferson County Colorado	527,056	473,495	90%	508,254	96%
Under 18	133,486	101,535	76%	108,618	81%
18 and over	393,570	371,894	94%	399,575	102%

Table 2. AREX Administrative Records Person Counts

	In Bottom-up	Not in Bottom-up	Total
In Top-down	2,275,456	70,031	2,345,487
Not in Top-down	7,525	---	---
Total	2,282,981	---	---

Table 3. Geographic Differences for Persons in both the Top-down and Bottom-up

Top-Down in Bottom-up	Same Address	Different Address	Different Block	Different Tract	Different County
2,275,456	2,258,441	17,015	15,129	11,847	2,363
100.0%	99.3%	0.8%	0.7%	0.5%	0.1%

Table 4. Comparison of Census and AREX household size, by NRFU status, and by imputation status—for linked housing units

AREX person count compared with Census	All Census housing units	Census non-NRFU housing units	Census NRFU housing Units	Non-imputed Census housing units	Imputed vacant Census housing units	Imputed occupied Census housing units
Same count	454,437 (51.1%)*	359818 (56.8%)	94,619 (37.0%)	449,582 (51.4%)	71 (26.5%)	4,784 (31.8%)
AREX one higher than	124,706 (14.0%)	84,269 (13.3%)	40,437 (15.8%)	122,519 (14.0%)	95 (35.5%)	2,092 (13.9%)
AREX one lower	127,531 (14.3%)	85,178 (13.4%)	42,353 (16.5%)	124,355 (14.2%)	0	3,176 (21.1%)
AREX 2 or 3 higher	64,635 (7.3%)	36769 (5.8%)	27,866 (10.9%)	63,024 (7.2%)	77 (28.7%)	1,534 (10.2%)
AREX 2 or 3 lower	79,848 (9.0%)	47,938 (7.6%)	31,910 (12.5%)	77,463 (8.9%)	0	2,385 (15.9%)
AREX 4 or more higher	15,781 (1.8%)	6,486 (1.0%)	9,295 (3.6%)	15,316 (1.8%)	25 (9.3%)	440 (2.9%)
AREX 4 or more lower	22,700 (2.6%)	13,158 (2.1%)	9,542 (3.7%)	22,068 (2.5%)	0	632 (4.2%)
Total	**889,638 (100%)**	**633,616 (100%)**	**256,022 (100%)**	**874,327 (100%)**	**268 (100%)**	**15,043 (100%)**

*** Percents are percents of column total**

Table 5. Comparison of AREX and Census demographic composition of households. For linked households with the same number of people only, by size

HH Size		Total	Equal for all sex groups[1, 2]	Equal for all race groups	Equal for all Hisp. groups	Equal for all 5-year age groups	Equal for age groups 0-17,18-64, 65+	Equal for all demographic groups[3]
All	NRFU	85,774	81.0%	87.7%	92.3%	58.1%	84.9%	63.4%
	non-NRFU	359,652	93.7%	94.7%	95.3%	86.9%	95.0%	84.6%
1	NRFU	31,313	82.5%	89.3%	95.7%	57.5%	91.1%	68.9%
	non-NRFU	107,979	95.0%	96.8%	98.1%	89.7%	97.5%	90.2%
2	NRFU	24,499	83.7%	88.5%	92.7%	58.6%	83.6%	64.9%
	non-NRFU	133,760	95.7%	96.0%	96.5%	88.6%	95.9%	87.9%
3	NRFU	12,549	75.7%	85.6%	89.4%	54.3%	77.1%	54.8%
	non-NRFU	48,092	90.1%	92.1%	93.0%	81.4%	91.4%	76.8%
4	NRFU	11,423	79.8%	86.3%	88.4%	63.2%	83.3%	60.2%
	non-NRFU	48,758	91.5%	91.7%	91.2%	84.9%	93.7%	77.3%
5	NRFU	4,473	78.1%	84.9%	87.2%	60.4%	80.0%	56.8%
	non-NRFU	16,250	89.2%	90.1%	89.9%	81.8%	91.4%	73.0%
6	NRFU	1,269	71.0%	80.4%	83.0%	54.0%	73.1%	46.8%
	non-NRFU	4,090	83.4%	87.8%	86.9%	72.4%	84.6%	63.0%
7+	NRFU	248	53.6%	79.8%	81.2%	27.0%	47.6%	24.6%
	non-NRFU	723	58.0%	81.2%	80.0%	29.3%	54.5%	30.2%

(Table 4 notes)

1. I.e., the AREX and Census households have the same number of males and the same number of females.
2. Percents are percents of Total.
3. Both sex groups, all race groups, both Hispanicity groups, and age groups 0-17, 18-64, 65+

Table 6. Address contains only persons 65 and older versus demographic match/non-match status

	All AREX persons age 65 or older?		
	No	**Yes**	**Total**
Non-match	513,926	33,418	547,344
	66.56%	28.43%	
Match	258,150	84,144	342,294
	33.44%	71.57%	
Total	772,076	117,562	889,638
	86.79%	13.21%	100%

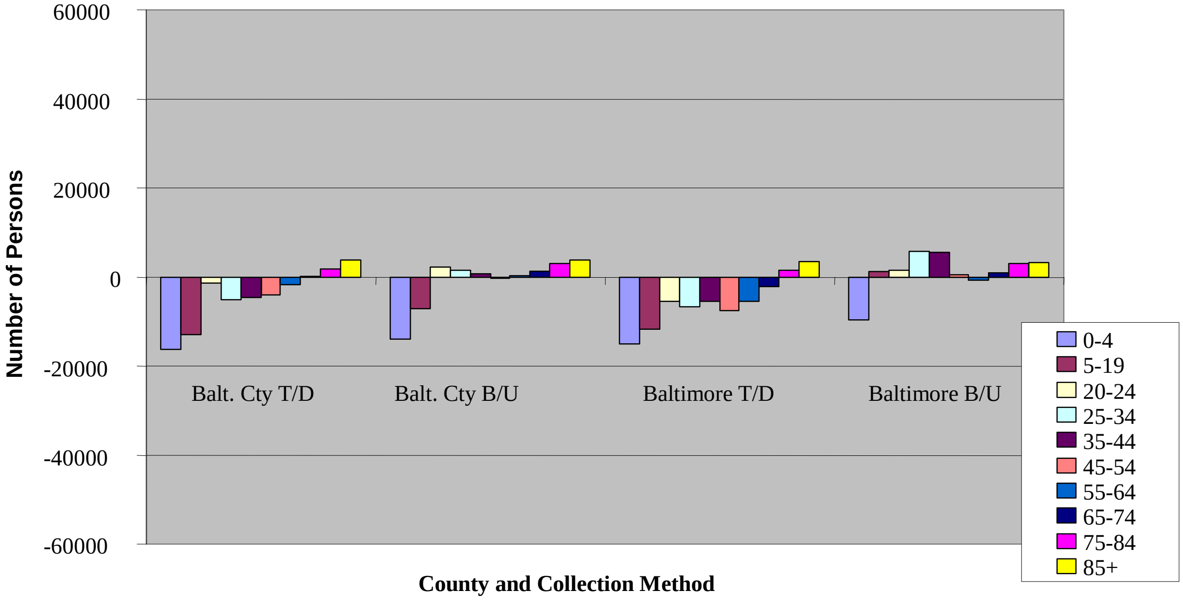

Figure 1: Net Population Difference by Age, County, and Top-down versus Bottom-up method—Baltimore City and Baltimore County. (Colorado results are similar.)

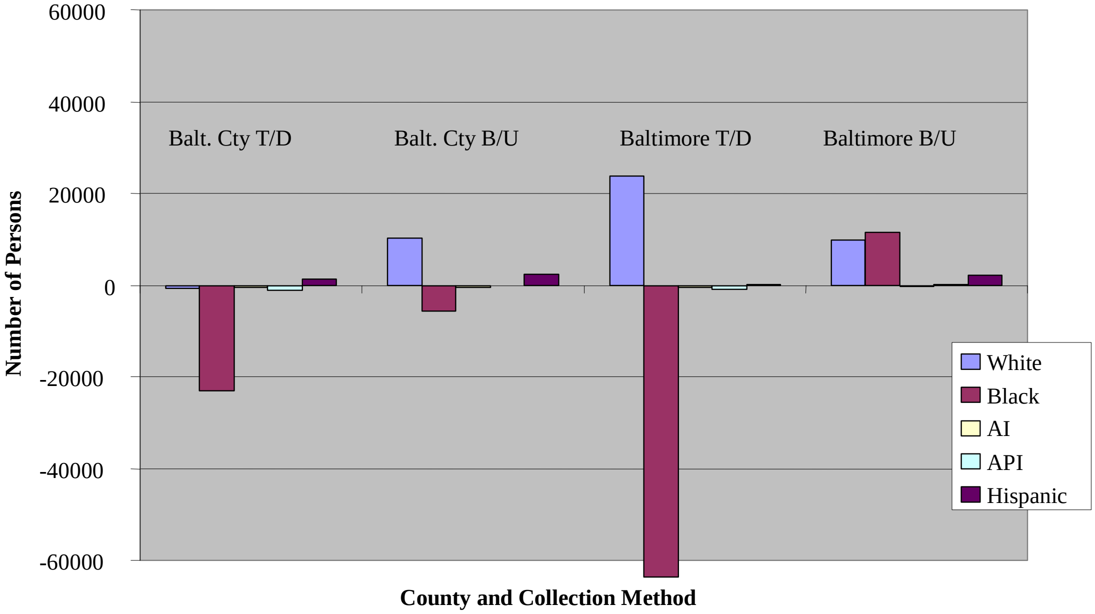

Figure 2: Net Population Difference by Race, County, and Top-down versus Bottom-up method—Baltimore City and Baltimore County. (Colorado results are similar.)