ADMINISTRATIVE RECORDS EXPERIMENT IN 2000: OUTCOMES EVALUATION

Harley K. Heimovitz
Planning, Research, and Evaluation Division, U.S. Census Bureau, Washington, DC 20233

Keywords: Administrative Records, Census, Enumerate, Evaluation, Imputation, Population

SUMMARY

This paper presents selected results from the Administrative Records Experiment in 2000 (AREX 2000). AREX 2000 used administrative records to enumerate the population in two test sites and compared the results to Census 2000. The test sites included two Maryland and three Colorado counties that offered distinct challenges to the enumeration process. The Outcomes Evaluation assessed two enumeration methods, compared county and sub-county population counts to Census results, and examined the impact of race imputation and other processing issues. The results confirm that administrative records provide good estimates of Census household population counts at larger geographies with greater accuracy using the Bottom-Up enumeration method. Both Top-Down and Bottom-Up methods undercounted Census in most counties, with Bottom-Up population counts ranging from 97-102% of Census results. Age and sex differences indicated problems with source administrative files. Most of the AREX race distributions did not accurately replicate Census results, which was attributed to weaknesses in the race imputation methodology. Imputation rates and type of imputation, housing unit characteristics, and presence of non-relative household members were all associated with AREX-Census differences.

BACKGROUND

In *Modernizing the U.S. Census*, the National Research Council evaluated the use of administrative records as a 'radical alternative' to a traditional census and small area population estimates (Edmonston and Schultze, 1995:167). They recommended that research proceed on both applications, noting that administrative records data are 'a major resource, both potential and realized, in the development and production of small area estimates.' This paper presents selected results from the AREX 2000 Outcomes Evaluation. AREX 2000 involved the construction of a test file to simulate Census 2000 results and evaluated the limitations of source files and enumeration methods. The AREX enumeration process included two sites, totaling about one million housing units and 2.6 million persons. The Maryland site included Baltimore City and County. The Colorado site included Douglas, El Paso, and Jefferson Counties. Selection of the five counties was based on whether housing units were assumed easy-to-enumerate in an administrative records census or hard-to-enumerate.[1] Baltimore, Douglas, and Jefferson Counties were considered easy-to-enumerate areas, while Baltimore City and El Paso County were hard-to-enumerate. The Outcomes Evaluation report emphasizes the AREX household population, while the AREX Process and Household Evaluation reports include group quarters. Group quarters may require special data acquisition and processing operations and have been excluded from most of the analyses in the Outcomes Evaluation.

Selection of administrative files

The file selection process was based on previous administrative records experiments and research on coverage overlap between the files (Huang and Kim, 2000; Prevost, 1997; Sweet, 1997). AREX 2000 relied upon national files, due to the limitations of state and local data. Differences in state measures, reporting timelines, and availability of files make it difficult to achieve consistent national coverage of many data items. Interagency contractual arrangements for using state and local files are intricate and time-consuming to develop. And linking commercial and government databases heightens public sensitivity to privacy issues (general privacy issues are discussed in Gellman, 1997). The files used in AREX 2000 include:

- Internal Revenue Service (IRS) Individual Master File (1040) for tax year 1998
- IRS Information Returns File (W-2/1099) for tax year 1998
- Department of Housing and Urban Development 1999 Tenant Rental Assistance Certifications System (TRACS) File
- Center for Medicare and Medicaid 1999 Enrollment Database
- Indian Health Services 1999 Patient Registration File
- Selective Service 1999 Registration File

File processing

The administrative files were assembled and unduplicated using Social Security Numbers (SSNs). Address-processing methods were then applied to the linked administrative files and persons were assigned block-level geographic codes and addresses. The two address-processing or enumeration methods are described as Top-Down and Bottom-Up:

Top-Down - The Top-Down method provides block-level counts of the AREX population. The administrative records files were unduplicated using SSNs and addresses

[1] Criteria included city-style address, single/multi-unit dwelling, householder age, and demographic characteristics.

and assigned to Census tabulation blocks using TIGER files. This method does not provide a census of households or housing units and includes both household and group quarters (GQ) residents.

Bottom-Up **-** The Bottom-Up approach unduplicated administrative records files using SSNs and addresses, and the results were matched to a list of residential addresses on the Census Master Address File (MAF). Non-matching AREX addresses were remedied by replacing them with corresponding Census records ('census pull' process). Conceptually, the Bottom-Up tally file is like an edited and enhanced version of the Top-Down tally file. However, the file processing procedures used two parallel and independent methods to create the Top-Down and Bottom-Up files.

METHODOLOGY

The Outcomes Evaluation measured how well AREX simulated Census 2000 results at county and subcounty levels and identified weaknesses in AREX processing. Differences between Bottom-Up and Top-Down enumeration results and key demographic characteristics were assessed. The evaluation used various methods to accomplish its objectives, including univariate and multivariate statistical analyses of AREX-Census differences, and spatial/ecological maps that examined the distributions of key measures. The Outcomes Evaluation tried to disentangle the influence of demographic change, AREX processing operations, and coverage and data quality issues. The key research question in this evaluation was:

What factors influenced the accuracy of the AREX county and subcounty results, what actions could improve the quality and coverage of administrative records, and what are the limitations of administrative records as a reliable source of intercensal population counts?

'Zero-blocks' occur when AREX reports persons having a particular characteristic but Census does not. Because Census was used as the standard and denominator for algebraic percent errors (ALPEs), these zero-blocks were undefined. However, county and tract-level population counts and comparisons included these blocks because they were aggregated at larger geographies. Inflated ALPEs occur because some blocks (or tracts) had very small denominators that tended to produce large ALPEs, despite small differences between AREX and Census counts.

The terms 'undercount' and 'overcount' describe how well AREX counts matched Census results and have no further connotation. That is, undercounts and overcounts reflect any of several problems, including coverage issues, demographic change, and processing errors. Some administrative records addresses did not match Census addresses and were replaced with census household data (census pull).

Variable definitions used in the analyses include:

Algebraic percent error (ALPE): AREX and Census counts were the inputs for calculating the algebraic percent difference with Census counts as the standard.

Race: Both AREX and Census versions of this variable used single race values with categories White, Black, American Indian, and Asian-Pacific Islander. The Hispanic origin of the race categories was ignored.

Hispanic origin: AREX and Census versions of this variable used Hispanic origin and ignored race category.

Population density: Population density of blocks and tracts was calculated using Census total population values.

Neighborhood characteristics: A classification of neighborhoods was constructed from block-level Census attributes that included population demographics, population density, and housing characteristics.

Vacancy rate: Vacancy rate used Census-reported values of housing unit vacancies within blocks and tracts.

Rental rate: Rental rate used Census-reported values of home tenure within blocks and tracts.

Non-relative household members: Census-reported values of housing units with non-relative household members.

Multivariate analyses

AREX-Census differences were examined in multinomial categorical regression models predicting block-level total population ALPEs. But the distribution of ALPEs is truncated at –1 when the AREX population equals zero, and small Census blocks have inflated overcounts. To compensate for this difficult to transform ALPE distribution, the values were categorized into groups. Each block-level ALPE was assigned to one of five subgroups based on their interquartile ranges. Groups 1 and 2 of the interquartile groups included undercounts, while Groups 4 and 5 were overcounts. Group 3 had the smallest ALPE scores (both under- and overcount) and included *real* zero-scores (*zero-blocks* were excluded). Categorical regression models (SAS PROC CATMOD) compared each of the four groups to the reference group using all blocks with complete data.

RESULTS

County-level results

The county-level analysis results are presented in Tables 1 and 2 and summarized below:

- Bottom-Up county ALPEs were smaller than Top-Down results; Bottom-Up ALPE improvements were variable: both Jefferson County and Baltimore City had Top-Down ALPEs of -8.8%, but Bottom-Up for Jefferson County was -3.0%, compared to +1.8% for Baltimore City.

- The smallest total population Bottom-Up ALPE was in Baltimore County (-1.1%); the largest Bottom-Up ALPE was in Douglas County (-3.2%).
- Male and female ALPEs were small in all five counties and ranged from –4.0% to +4.2%.
- Generally, younger age groups had the largest negative ALPEs in all five counties; age 0-4 ALPEs ranged from –33.8% in Jefferson County to –23.4% in Baltimore City.
- Older age groups tended to have positive ALPEs that increased for older age groups.
- Blacks were overcounted in all three CO counties and Baltimore City and undercounted in Baltimore County where Blacks are the largest minority group.
- Hispanics were overcounted in both MD counties and undercounted in all three CO counties where Hispanics are the largest minority group.
- American Indians had the greatest ALPEs in all five counties; ALPEs ranged from –34.1% in Jefferson County to –11.3% in Baltimore City.
- Asian-Pacific Islanders were overcounted in all three CO counties and Baltimore City and undercounted in Baltimore County.

Bottom-Up ALPEs were generally smaller due to more stringent address-matching requirements (compared to Top-Down). The census pull replacement of unmatched AREX addresses also reduced differences. The overall impact of the Bottom-Up method was to increase the number of AREX households and eliminate unverified households that place persons in the wrong blocks. Ideally, this process increases the accuracy of demographic characteristics in small areas.

Female undercounts were slightly worse than male undercounts. Some women may be less active within the administrative records systems. For example, studies indicate that lifetime labor force participation varies by a woman's race/ethnicity, health status, and caregiving experiences (Flippen and Tienda, 2000). Differential male-female undercounts may also be due to delayed reporting of mortality because men and women have different survival rates at older ages. Lagged reporting and differential mortality produce overcounts that appear to offset male undercounts.

Age ALPEs were large due to the combined effect of errors in the administrative record collection process and recording lag from demographic processes. Infants are likely to have poor coverage due to administrative delays in reporting their births. Households with five or more children, new dependents born between the beginning of tax year 1999 and the April 1, 2000 date of the Census, and separated or remarried parents who did not claim a child in their tax return are also likely to have incomplete coverage of household members. This was demonstrated by the large undercounts for the 0-4 age group. College-aged persons may have been reported at a parent's IRS tax address but actually reside on a campus in a different area. The 20-24 year age group also had large ALPE overcounts in some of the AREX counties. Persons aged 65+ were generally overcounted in all five counties, which may be due to administrative records not capturing migration (to new residences and nursing homes) and mortality of older persons. Despite linkages to Medicare records, some older persons (age 65+) may have less reliable information in administrative records because lagged reporting may count persons alive and resident who may have died or moved.

An important difference between the Top-Down and Bottom-Up results was the manner in which the race imputation model treated children. The Top-Down method did not impute the race of children. In the Bottom-Up process, children were assigned the race of the primary tax filer at their address. The 1998 tax returns linked the householder and first four dependents, allowing householder race to be assigned to dependents. For traditional married families, it is likely that only three children plus the spouse were linked to the householder. While a formal evaluation of the revised race imputation methodology has not been conducted, it is assumed that the more stringent Bottom-Up address requirements and use of tax filer race improved the accuracy of race assignment for children.

Whites and Blacks were overcounted in four of the five counties (Bottom-Up results), but were generally *undercounted* in the Top-Down results (not shown). These differences provide further support that the assignment of White and Black race codes was imprecise, due to deficiencies in the race imputation model. The race imputation model also exhibited 'regression towards the mean' in assigning Black and White races, because aggregate population estimates were used to estimate individual race characteristics.

American Indians had large undercounts in all counties due to the small population bases used in the ALPE calculations. Hispanics had large undercounts in all three CO counties and overcounts in the MD counties. Hispanics are a larger percentage of the total population in CO (5-11%) but a smaller percentage and number of the MD population (less than 2%). The results suggest that the undercount may be due to problems with race coding, the race imputation model, recent Hispanic migrants to CO not recorded in AREX, or persons not appearing in administrative records. For example, casual labor and domestic workers may receive cash payment, provide false SSNs, and may not exist in administrative records. That is, they appear in Census, but migration, type of employment, and AREX processing limitations may be associated with their undercounting. Asian-Pacific Islanders were undercounted in Baltimore County, with large overcounts in El Paso and Jefferson Counties and smaller undercounts in the remaining counties.

Tract and block-level analyses (tables not shown)

The tract- and block-level ALPE results describe the accuracy of counts at the smallest geographic levels. One problem with this type of comparison is the ALPE denominator potentially inflates block-level ALPEs for small population subgroups and especially minorities. This inflation is likely to be greater than found in the county comparisons. A second issue is the exclusion of blocks where Census did not identify persons with a particular attribute (zero-blocks). Tract and county ALPEs include blocks with zero counts because these blocks were included in larger geographies. However, the block-level ALPEs use the reduced sample of blocks and the results may be quite different when comparing the ALPEs at various geographic levels:

- More than 75% of tracts had AREX total population counts within +/-5% of Census results (5% criterion), and more than 95% of tracts had counts within 25% of Census (25% criterion) in four of five counties; Baltimore City had less accurate results with about 50% of tracts exceeding +/-5% of Census results.
- A larger proportion of tracts had moderate and large ALPE undercounts, compared to overcounts.
- AREX was less accurate in estimating blocks than tracts in all counties; from 18-39% of blocks were within the 5% criterion, and about 85% were within the 25% criteria in the five counties; Douglas County had the best results at the 5% criterion and Baltimore County was best at the 25% criterion.
- In the MD counties, slightly more blocks had moderate or large overcounts (ALPEs exceeding 5%), compared to the CO counties where more blocks had moderate undercounts (-5% to -24%).

Though the tract-level ALPEs for the total population resemble county-level results, the distributions indicate more Baltimore City tracts were overcounted. It's unclear whether these overcounts are related to persons who were actually uncounted in Census, or more likely, flaws in AREX processing. Households may have been added through the census pull process that replaced unmatched addresses that existed in other tracts or addresses. The AREX counts were less accurate at the tract- and block-levels due to incorrect assignment of households that average out for county-level counts. This was demonstrated by the greater number of moderate and large ALPEs, a consequence of 'regression towards the mean,' smaller denominators inflating ALPEs, and unmeasured AREX processing flaws. Though zero-blocks were excluded and fewer blocks met the 5% criterion, a surprisingly large proportion of blocks met the 25% criterion in all five counties.

AREX processing and operational issues

Race assignment was based on three methods:

- Most frequent report from source administrative files.
- Imputed from race probability estimates and assigned to adults.
- Imputed from householder's race and assigned to children under 18 years old (Bottom-Up only).

Table 3 provides a summary of race imputation and census pull proportions. The imputed race assignments may increase AREX-Census differences while the census pull process improves the apparent accuracy of AREX. The distribution of imputed and census pull cases fell into several distinct patterns and later analyses showed how the race assignment process affected ALPE results:

- Race imputation was greater in the CO counties, especially for Whites and Blacks.
- Both MD counties had similar imputation rates, with the rate of census pull much greater for Baltimore City.
- Douglas County had large imputation rates for total population, most of the race categories, and a large census pull rate.
- Census pull rates were large for Baltimore City and Douglas County (see Table 3), but both counties also experienced significant population change between 1990 and 2000.

MULTIVARIATE ANALYSIS

The primary goal of the categorical regression models was to identify the key predictors associated with block-level under- and overcounts and account for differences between counties and AREX sites. The model results identify the key predictors of total population ALPEs and assume the Census results to be the 'truth' about the AREX population. The extensive univariate and bivariate analyses are confounded by the characteristics of blocks, tracts, and counties; that is, demographic, ecological, and socioeconomic characteristics. The multivariate models remove this confounding so that fair comparisons can be made between predictor variables and block-level ALPEs.

Categorical model results

The model results in Table 4 focus on predictors common to both sites. The models compared blocks with moderate and large under- and overcounts to a reference group whose AREX blocks counts were closest to Census results (the 'best' or reference group). Reference characteristics had the smallest ALPE results and reflected blocks with: low mobility rates (low vacancy, rental, non-relatives), low imputation and census pull rates, suburban or moderate population density, moderate White population proportions, no mention of Blacks or Hispanics, and a large proportion of persons aged 45-64.

Mobile population groups and multi-unit dwellings in urban areas were associated with undercounts for both sites. Family formation and the career pathways of persons in their early 20s are both important for local planning purposes but were undercounted by AREX. Presence of non-relatives suggests cohabiting partners who could also be in the family formation process.

AREX overcounts were unrelated to factors affecting undercounts. Vacant housing units in blocks with a lower population density suggest rural areas where there was net out-migration and/or housing unit turnover was slow. An address identified in AREX may have been vacated but there was no way to know this from the administrative file sources.

Generally, post-processing operations were associated with both under- and overcounts while the census pull operation substituted Census results for unmatched addresses. Post-processing was also associated with race ALPEs, consistent with other models showing less accurate results for race composition. The AREX 2000 Outcomes Evaluation has the complete results of the multivariate analyses.

CONCLUSION

This paper describes selected results from the AREX 2000 Outcomes Evaluation. The study compared AREX and Census 2000 population counts, assessed the strengths and weaknesses of administrative data as a supplement or substitute for Census, and compared address-matching and race imputation methods. Some general themes and their implications are summarized below:

1. AREX provided county-level population counts that were close to Census 2000 counts, though the results were not as good for subcounty and demographic counts.
2. The Bottom-Up enumeration method performed better than the Top-Down method, but had additional processing constraints.
3. The census pull process resolved some of the deficiencies in the address-matching process and may be useful in ongoing processing cycles.
4. The race imputation process did not perform well and requires improvement for subsequent processing.
5. Demographic events and/or reporting lag impacted the accuracy of AREX counts. Administrative records processing needs to synchronize dates in administrative data to replicate Census place-time reporting requirements, perhaps incorporating quarterly updates from data providers.
6. AREX counts for the oldest and youngest persons suggest that birth and death information was not recorded in a timely manner. Further research is needed to understand whether this was due to the agency providing the data or delays prior to their receipt (i.e., other agencies, their processing schedules, and state regulations and policies).
7. There was some suggestion that college-aged persons were not counted accurately. Migration was the likely determinant and research is needed that distinguishes temporary and permanent addresses of transient persons, as well as non-relative household members.
8. The model results reinforce the univariate and bivariate results (not shown in this report): vacancy and rental rates, presence of non-relatives, and race imputation were all associated with AREX-Census differences.

REFERENCES

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

Czajka, J.L., Moreno, L., and Schirm, A.L. (1997). *On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population*. Washington, DC: Internal Revenue Service.

Edmonston, B. and Schultze, C. (1995). *Modernizing the U.S. Census*, National Academy Press, Washington, DC.

Flippen, C. and Tienda, M. (2000). Pathways to Retirement: Patterns of labor Force Participation and Labor Market Exit Among the Pre-Retirement Population by Race, Hispanic Origin, and Sex. *Journal of Gerontology: Social Sciences*, 55B: 1, S14-S27.

Gellman, R. (1997). *Report on the Census Bureau Privacy Panel Discussion*. Unpublished document available from the U.S. Census Bureau, June 20, 1997.

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

Leggieri, C., Pistiner, A. and Farber, J.E.. 2002. Methods for Conducting an Administrative Census Experiment in 2000. *Proceedings of the American Statistical Association.*

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

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

Taueber, C., Lane, J., and Stevens, D. (2000). *The Why, What, and How of Converting Program Records and Summarized Survey Data to State and Community Information Systems*. Conference Paper presented at Developing Public Policy Applications with Summarized Survey Data and Community Administrative Records. Baltimore, MD, June, 2000.

Table 1: Top-Down and Bottom-Up Counts of Total Household Population by County[1]

	Top-Down Results				Bottom-Up Results			
	AREX	Census	Difference	ALPE	AREX	Census	Difference	ALPE
Baltimore County	696,183	736,652	-40,469	-5.5%	728,205	736,652	-8,447	-1.1%
Baltimore City	570,648	625,401	-54,753	-8.8%	636,729	625,401	+11,328	+1.8%
Douglas County	148,270	175,300	-27,030	-15.4%	169,640	175,300	-5,660	-3.2%
El Paso County	456,891	501,533	-44,642	-8.9%	494,253	501,533	-7,280	-1.5%
Jefferson County	473,495	519,326	-45,831	-8.8%	503,622	519,326	-15,704	-3.0%

[1] AREX Top-Down counts include persons later identified in Bottom-Up as group quarters residents; Bottom-Up Census comparison excludes group quarters residents and population counts may differ from the Process and Households reports.

Table 2: Bottom-Up ALPE Results by Demographic Characteristics

	Baltimore County	Baltimore City	Douglas County	El Paso County	Jefferson County
Total	-1.1%	1.8%	-3.2%	-1.5%	-3.0%
White	1.9%	5.0%	-1.0%	5.4%	1.1%
Black	-3.9%	2.8%	25.1%	11.3%	30.8%
American Indian	-24.0%	-11.3%	-21.4%	-20.5%	-34.1%
Asian Pacific Islander	-0.2%	1.4%	4.6%	18.1%	6.3%
Hispanic	17.1%	19.5%	-2.1%	-10.9%	-11.4%
Age 0-4	-30.8%	-23.4%	-29.2%	-30.1%	-33.8%
5-19	-4.9%	0.9%	-7.0%	-6.2%	-10.1%
20-24	5.0%	3.7%	48.6%	12.2%	15.5%
25-34	1.6%	6.4%	-1.2%	3.2%	1.1%
35-44	0.5%	5.6%	-1.6%	1.8%	-1.8%
45-54	-0.3%	0.7%	-0.8%	0.2%	-0.6%
55-64	0.4%	-1.2%	0.4%	0.6%	0.5%
65-74	2.6%	2.0%	2.0%	2.9%	1.9%
75-84	7.6%	9.9%	1.0%	3.6%	2.5%
85+	38.2%	37.6%	77.1%	37.4%	35.1%
Male	-0.4%	4.2%	-2.5%	-0.7%	-2.5%
Female	-1.9%	-0.3%	-4.0%	-2.3%	-3.6%

Table 3: Race and Ethnicity Imputation Rates by County[1]

	Baltimore County	Baltimore City	Douglas County	El Paso County	Jefferson County
All persons	12.5%	9.8%	17.1%	16.9%	17.4%
White	12.0%	11.0%	16.8%	16.9%	15.6%
Black	11.7%	8.9%	26.7%	15.3%	31.6%
American Indian	28.8%	24.2%	26.3%	20.2%	21.1%
Asian Pacific Islander	28.7%	20.7%	28.2%	27.4%	31.2%
Race Unknown	0.2%	0.1%	0.2%	0.8%	0.6%
Hispanic	92.5%	82.6%	84.6%	85.3%	88.2%
Census Pull	6.3%	15.3%	13.5%	9.3%	7.4%

[1] (Imputed PCF + householder-assigned records to children) / total AREX records; Bottom-Up results.

Table 4: Key Predictors from Categorical Regression Models[1]

Outcome	Key predictors for both sites
Large undercounts (~ 15+%)	rental units, nonrelatives in HH, high population density, ages 0-4 and 20-24
Moderate undercounts (~ 5-15%)	rental units, nonrelatives in HH, AREX post-processing, ages 0-24
Moderate overcounts (~ 5-18%)	vacant units, AREX post-processing
Large overcounts (~ 18+%)	vacant units, rental units, low population density, AREX post-processing

[1] Relative to reference category of +/-5% of block-level Census counts; Bottom-Up results.