EVALUATING THE ABILITY OF ADMINISTRATIVE RECORDS DATABASES TO REPLICATE CENSUS 2000 RESULTS AT THE HOUSEHOLD LEVEL[1]

D.H. Judson and Mark Bauder, Bureau of the Census
4700 Silver Hill Rd, Suitland, MD 20746

KEY WORDS: Administrative records experiment, nonresponse followup, logistic regression

1. Introduction

The Administrative Records Experiment 2000 (AREX 2000) was an experiment in two areas of the country designed to gain information regarding the feasibility of conducting an administrative records census (ARC) or the use of administrative records in support of conventional decennial census processes. The first experiment of its kind, AREX 2000 was part of the Census 2000 Testing, Experimentation and Evaluation Program. The results of the testing will lead to recommendations for subsequent testing and ultimately for the design of the next decennial census.

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. Sailer, Weber, and Yau (1993) noted that counts of IRS person records, when properly corrected for coverage, were notably concordant with U.S. population estimates. There have been a number of other calls for ARC research--see for example Myrskyla 1991; Myrsklya, Taeuber and Knott 1996; Czajka, Moreno and Shirm 1997; Bye 1997.

More recently, direct use of administrative records in support of decennial applications was cited in several proposals during the Census 2000 debates on sampling for nonresponse followup (NRFU). The proposals ranged 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). AREX 2000 provided the opportunity to explore the possibility of NRFU support.

Demographically, the AREX provided date of birth, race, Hispanic origin, and sex, although the latter is not required for apportionment or redistricting purposes. Geographically, the AREX operated at the level of basic street address and corresponding Census block code. Unit numbers for multi-unit dwellings were used in address matching operations. In addition, the design did not provide for the collection of sample long form population or housing data, needs that will presumably be met in the future by the American Community Survey program. The design did assume the existence of a Master Address File and geographic coding capability similar to that available for the 2000 Census.

2. Methodology

The general goal of this evaluation is to focus on household-level comparisons. In the process, we will examine several difficult to measure aspects of the enumeration process: Nonresponse follow-up (NRFU) households, and households for which occupancy status and household demographics were wholly imputed ("unclassified" households). We will specifically assess the ability of AREX databases to match the demographic distributions of all households, NRFU households, and unclassified households. Finally, we will attempt to assess our ability to predict when an AREX household is likely to demographically match a census household.

One of the most important potential uses of administrative records data is to substitute administrative records data for some proportion of the nonresponse followup universe, or for the unclassified universe. In order to effectively use administrative records databases for substitution purposes, we must determine which kinds of administrative record households are most likely to yield similar demographic distributions to their corresponding census households. The purpose of the prediction section is to make this evaluation.

We refer to a pair of addresses (AREX and Census) that were linked through a computerized record linkage process as "linked" housing units. We use the term "imputed household" for unclassified addresses whose occupancy status and household characteristics have been imputed. We use the term "demographic match" when two households have the same age, race, sex and Hispanic origin (ARSH) distribution.

3. Limitations

Several individual limitations of the files themselves are worthy of note. First, AREX 2000 used files that were a year or more older than the target date of Census day.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

This means that movers, births, deaths, immigration and emigration, new housing, abandoned and demolished housing are unaccounted for. Second, AREX 2000 has difficulty enumerating children properly, by virtue of the time lag problem and by virtue of the limited demographics available for children on the Numident file (a source file for individual demographics; Miller, Judson, and Sater, 2000). Third, the race measurement and reporting deficiencies of the AREX 2000 experiment cause comparisons by race and Hispanic origin to be more challenging. In particular, most persons of Hispanic origin were imputed as such by AREX, thus complicating comparisons. Of course, Census 2000 multiple race reporting additionally complicates comparisons between AREX and Census households.

4. Descriptive Results

What are the basic characteristics of Census address data?

In the five counties covered by the AREX experiment, Census contains 1,092,460 housing units (HUs) and 1744 group quarters (GQs). Because AREX contains no administrative records data for Census GQs, we do not include Census GQs in these analyses. About 24,584 (2.3%) of Census households are "imputed households." About 360,914 (33.0%) are in the Census NRFU universe.

What are the basic characteristics of AREX address data?

As part of the implementation of the Bottom-Up method here, Census data were included in the AREX results for Census addresses with which no administrative records could be linked. We do not include them in the analyses, because we want to analyze the coverage and accuracy of administrative records data. There are 1,065,031 remaining AREX addresses.

Of these 1,065,031 AREX addresses, 56,638 were not linked with any Census address. 1,008,393 were linked with Master Address File (MAF) addresses. Because of the version of the MAF we used for matching addresses, some of those addresses did not exist on the final Census 2000 file. There were 992,865 AREX address that were linked with addresses that existed in the final Census. Of those that were linked with Census addresses, 889,638 are “one to one links." These are linked AREX-Census address pairs in which each address was linked with *exactly* one address. There were also "many to one" or “many to many” links - both where an AREX address was linked with more than one DMAF address, and where more than one AREX address was linked with one DMAF address. In what follows, "linked" addresses are always those that were “one to one” matches.

Tables 1-5 provide descriptive results on linked addresses. In particular, the link rate averaged 81.4% overall, and this varied by whether the address was vacant or not (Table 1). In particular, the difference between NRFU and non-NRFU link rates (70.9 and 88.4, respectively) can largely be explained by the fact that NRFU addresses are more likely to be vacant (Table 2). The effect was similar, although not as pronounced, for imputed addresses (Table 3). For linked addresses, AREX and Census obtained the same number of persons in the address 51.1% of the time, and plus or minus one person 69.4% of the time (Table 4).

Finally, of the linked addresses that had the same number of people, the AREX demographic composition matched the Census demographic composition 80.5% of the time and this tended to go down as the Census household size went up (Table 5).

5. Predictive Results

The purpose of the predictive analysis is to ascertain what individual variables, obtainable before census operations, would allow us to predict when an an AREX address would demographically match its census counterpart. If we can predict with reasonable accuracy, then we can potentially use administrative records data to “substitute” for census data in non-response followup or imputation. This analysis is primarily exploratory.

Tables 6-8 illustrated preliminary tabulations for variable selection. In these tables the gray shaded percentages are compared, and column variables with a large *difference of proportions* are deemed discriminative and entered into a logistic regression equation. Three large effects are described: Single-unit versus multi-unit addresses (30.52% matched versus 40.67%, Table 6); number of persons in the address (one or two person households matched about 50%, Table 7); and when the AREX address contained only persons 65 and older (71.57% matched versus 33.44%, Table 8).

Finally, Table 9 presents selected estimated odds ratios for selected variables in the model, for main effects and interaction effects. They can be interpreted as the expected change in the odds of a household demographically matching for the variable going from zero to one, holding other effects constant.

6. Conclusions and Recommendations

We can summarize our conclusions thus: 81.4% of the Census addresses (computer) linked on a one-to-one basis with an AREX address. Within these linked addresses, AREX and Census match the number of of persons 51% of the time. Of that 51%, demographics match 80% of the time. When we compare AREX and NRFU households, the demographic matches are less likely to occur, but this leads to the question: Is the poor match a result of poor quality AREX data or poor quality NRFU data? Or, more broadly, is the vintage of the AREX files to blame, with various demographic events being poorly captured? Finally, computer record linkage error could have created links that are false, contributing to demographic nnonmatches. Finally, we have developed a moderately predictive model that allows us to predict when an AREX

address will demographically match a Census address. We recommend that the Census Bureau continue to improve computerized record linkage; develop methods to reduce the time lag of AR data; test AR data for NRFU substitution and imputation purposes in future census tests; test AR data for MAF improvement; continue to improve race and Hispanicity modeling and imputation; and continue to explore uses of modeling for predictive or calibration purposes.

Table 1. Coverage by AREX of Census housing units.

	Total	Linked with AREX housing units (% of total)	Linked with AREX occupied housing units (% of total)	Linked with AREX vacant housing units (% of total)
Census housing units	1,092,460	889,638 (81.4%)	813,688 (74.5%)	75,950 (7.0%)
Occupied Census housing units	1,017,273	854,741 (84.0%)	787,802 (77.4%)	66,939 (6.6%)
Vacant Census housing units	75,187	34,897 (46.4%)	25,886 (34.4%)	9,011 (12.0%)

Table 2. Coverage by AREX of Census housing units, by NRFU status.*

Type of Census housing unit	Total	Linked with AREX housing units	Linked with AREX occupied housing units	Linked with AREX vacant housing units
NRFU	360914	70.9%	60.8%	10.1%
non-NRFU	716450	88.4%	82.9%	5.5%
Occupied NRFU	289224	76.7%	67.1%	9.6%
Occupied non-NRFU	715115	88.5%	83.0%	5.5%
Vacant NRFU	71690	47.6%	35.2%	12.3%
Vacant non-NRFU	1335	58.7%	46.3%	12.4%

* This analysis does not include 15,096 housing units in Census whose NRFU status is not indicated in the file.

Table 3. Coverage by AREX of Census housing units, by imputation status.

Type of Census housing unit	Total	Linked with AREX housing units	Linked with occupied AREX housing units	Linked with vacant AREX housing units
Imputed	24,584	62.3%	51.7%	10.5%
Non-imputed	1,067,876	81.9%	75.0%	6.9%
Imputed occupied	23,811	63.2%	52.6%	10.6%
Non-imputed, occupied	993,462	84.5%	78.0%	6.5%
Imputed vacant	773	34.7%	25.5%	9.2%
Non-imputed, vacant	74,414	46.5%	34.5%	12.0%

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%)	94619 (37.0%)	449,582 (51.4%)	71 (26.5%)	4,784 (31.8%)
AREX one higher than	124,706 (14.0%)	84269 (13.3%)	40437 (15.8%)	122,519 (14.0%)	95 (35.5%)	2,092 (13.9%)
AREX one lower	127,531 (14.3%)	85178 (13.4%)	42353 (16.5%)	124,355 (14.2%)	0	3,176 (21.1%)
Several	rows	omitted				
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. Comparisons between AREX and Census for demographic groups, for linked households with the same number of people only.

HH Size	Total linked, of equal size	Equal for all sex groups [1]	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 sizes	445,426	91.2%[2]	93.4%	94.8%	81.3%	93.1%	80.5%
1	139,292	92.2%	95.1%	97.5%	82.5%	96.1%	85.4%
2	158,259	93.8%	94.8%	95.9%	83.9%	94.0%	84.3%
3	60,641	87.1%	90.7%	92.3%	75.7%	88.4%	72.2%
4	60,181	89.3%	90.7%	90.7%	80.8%	91.7%	74.0%
5	20,723	86.8%	88.9%	89.3%	77.2%	89.0%	69.5%
6	5,359	80.4%	86.0%	86.0%	68.0%	81.8%	59.2%
7+	971	56.8%	80.8%	83.0%	28.7%	52.7%	28.7%

[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 the Total column**
[3] **Both sex groups, all race groups, both Hispanicity groups, and age groups 0-17, 18-64, 65+**

Table 6. Single unit or multi unit address (from AREX) and demographic match/nonmatch status.

	From AREX data: Single unit at BSA	Multi unit at BSA	Total
Not Matched	413,638 59.33	133,706 69.48	547,344
Matched	283,566 40.67	58,728 30.52	342,294
Total	697,204 78.37	192,434 21.63	889,638 100

Table 7. Number of persons in the AREX address versus demographic match/nonmatch status.

	AREX household number of persons 0	1	2	3	4	5	6+	Total
Not Matched	66,939 88.14	106,529 49.27	119,419 48.88	105,101 72.2	78,193 65.19	39,826 74.84	31,337 91.03	547,344
Matched	9,011 11.86	109,680 50.73	124,895 51.12	40,475 27.8	41,756 34.81	13,390 25.16	3,087 8.97	342,294
Total	75,950	216,209	244,314	145,576	119,949	53,216	34,424	889,638

Table 8. All AREX persons age 65 or older versus demographic match/nonmatch status.

	All AREX persons age 65 or older? No	Yes	Total
Not Matched	513,926 66.56	33,418 28.43	547,344
Matched	258,150 33.44	84,144 71.57	342,294
Total	772,076 86.79	117,562 13.21	889,638 100

Table 9. Selected estimated odds ratios from logistic regression model.

Selected main effects	
Non-multi unit:	2.6
One or two persons in HH:	3.5
No AREX imputed race:	2.1
AREX one or more white:	2.1
All AREX 65 and older:	1.7
Selected interaction effects:	
Total effect of 65+,nonmulti,nonimputed:	5.2
Total effect of 65+,1+white, 1-2 persons:	19.2

7. References

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

Bye, Barry (1997). "Administrative Record Census for 2010 Design Proposal." Washington, D.C.: United States Department of Commerce.

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, D.C.: Internal Revenue Service.

Edmonston, Barry and Schultze, Charles (1995) *Modernizing the U.S. Census*, National Academy Press, Washington D.C.

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

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.

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

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

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

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

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