Identifying Residents in Administrative Records That Do Not Match the Census

Michael D. Larsen[1], The University of Chicago, Department of Statistics

In Census 2000, the United States Bureau of the Census will try to count everyone living in the United States on Census Day, April 1, 2000, at their residences according to Census Residency rules. In 1998, the Bureau of the Census tested new methodology and processes in the Census 2000 Dress Rehearsal. Most residents were enumerated at the their residences, but some were missed.

Administrative records were collected, unduplicated, and matched to Dress Rehearsal census records. Some administrative records that did not match census records corresponded to people who should have been enumerated on Census Day, whereas some represented people who were not residents, perhaps because they had moved from the area. The residence status of a sample of the administrative records in each site was determined by field interviews.

Simply adding unmatched administrative records to the census counts would introduce many errors, because many are nonresidents. Searching for all individuals corresponding to administrative records is time-consuming and expensive and finds many nonresidents. Here, we report on criteria that could have been used to identify administrative records in the followup sample that corresponded to people who should have been enumerated. Residency status is known and will be used to judge how the criteria perform.

The second section describes the sites and record sources in the Dress Rehearsal. The third section describes methods used to identify groups of administrative records that contain higher than average proportions of residents. The fourth section discusses limitations of this research. The fifth section presents results for the three sites. The sixth section concludes and makes recommendations for Census 2000, ongoing research, and the 2010 Decennial Census. The seventh and eighth sections contain acknowledgments and references, respectively.

■ Census 2000 Dress Rehearsal

The Census 2000 Dress Rehearsal was conducted in 1998 in three sites: Sacramento, CA (an urban, primary media market with relatively high proportions of African-Americans, Hispanics, and Asian and Pacific Islanders); Columbia, SC and eleven surrounding counties (non-urban with a mix of city-style and other address types, as well as a relatively high percentage of African Americans); and Menominee, WI (a rural county with an American Indian Reservation). The three sites are very different in important characteristics, such as population size, number and median value of housing units, percent of housing units occupied and owner-occupied, poverty rate, vacancy rate, and percent movers. (Bureau of the Census, 1999a and 1999b; Prewitt, Shapiro, and Barron, 1999).

Mailout/Mailback, mailing forms to households that then return them via mail, was used in Sacramento and South Carolina. Update/Leave/Mailback, having census personnel visit addresses and locations, update address lists, and leave forms for residents to complete and mail back, was used in Menominee and in areas without city-style addresses in South Carolina. Non-Response Followup (NRFU), sending enumerators to addresses from which a form was not returned via mail, was done on a sample basis in Sacramento (Sample NRFU), but 100 percent in the other two sites. In Sacramento, an integrated coverage measurement (ICM) survey was used to produce a dual-system estimate (DSE) of the population count. In South Carolina, a Post-Enumeration Survey (PES) was undertaken to study population coverage in the traditional census. In Menominee, an ICM survey was used to improve coverage. The sites varied greatly in percentage population increase since 1990, size, and vacancy and mail response rates in the Dress Rehearsal. Prewitt, Shapiro, and Barron (1999) and Thompson (1999) note that plans for Census 2000 and the Dress Rehearsal are not identical.

The administrative records available at all three sites

were the Internal Revenue Service (IRS) Tax Year 1996 1040 Individual Master Tax Return File and the Selective Service System (SSS) registration file. Housing and Urban Development (HUD) Tenant Rental Assistance Certification System (TRACS) files were available in California and South Carolina. Department of Health and Human Services, Public Health Service Indian Health Service (IHS) patient registration files were available in California and Wisconsin. In South Carolina, State Medicaid enrollment data also were available. No other State or local files were available for this research; the decision to conduct this research was not made in time to obtain the files, and experience in 1995 and 1996 has shown the difficulty of dealing with these numerous files and their various contents, formats, and acquisition processes (Neugebauer, 1996; Neugebauer, Perkins, and Whitford, 1996; White and Rust, 1997). When a unique Social Security Number (SSN) was available on an administrative record and basic personal information was in correspondence, supplemental demographic information was found on the Social Security Administration (SSA)'s Numerical Identification (Numident) file.

Administrative records were linked to records on the Census Unedited File (CUF) by Master Address File Identification Number (MAFID) (see, e.g., Bureau of Census 1999b). Each MAFID corresponds to a housing unit and is geographically coded to the Topologically Integrated Geographic Encoding and Referencing (TIGER) system. Housing units with occupants are households. The various administrative records were unduplicated to produce a single record for each person with fields of information (e.g., name, race, gender) being completed, based on entries in administrative records selected according to a priority order for each field. For example, the order of files used to produce gender was first SSA, then Selective Service, Medicaid, HUD, IHS, and finally IRS. Post office box addresses, group quarters addresses, addresses without a block code, and addresses in ICM and PES census blocks were removed from consideration.

Nonmatching administrative records were stratified into nine strata, which were defined by crossing two characteristics of the records at a MAFID: the number of administrative sources for the housing unit and the number of matches to CUF in the housing unit. The sample of records selected for field followup to determine Census Day residency status was called the Coverage Improvement Followup (CIFU) study.

■ Methods

Initially, subgroups and combinations of subgroups of unmatched administrative records are identified that have relatively more residents than average. The records in the best subgroups could be chosen for review in future operations. Larsen (1999a) proposed an accumulation criterion that increases with the number of nonresidents, but decreases by a scalar times the number of residents in the subgroup and a simulation procedure to assess the number of groups to combine into a class of likely residents. Use of the criterion demonstrated that, in samples of the size in this study, usually, few groups should be combined.

Logistic regression produces equations to predict Census Day residency. Predictor variables are indicators of administrative source, dummy variables that indicate subgroups of records, and quantitative variables from the administrative records. The estimated equations are used to form groups of records that are predicted to be residents. Fivefold cross-validation (fitting the logistic regressions to four-fifths of the data, then predicting residency status for the remaining one-fifth of the data) is used to assess prediction error. Cross-validation resembles the application of an estimated prediction rule to data from a new, but similar site.

Other methods could be used to produce complex criteria for searching for residents, but are not explored here. In previous work, alternative methods made little difference in quality of final results (Larsen, 1999a).

Finally, county-level covariates from the Planning Data Base (Bruce and Robinson, 1999a, 1999b) in South Carolina are used to predict the residency rates among nonmatched administrative records by county. In future sites, prediction equations like those produced here could potentially identify areas in which nonmatching administrative records are more likely than average to contain residents. Linear regression is used for model-

ing, and diagnostic residual plots are discussed. There are few degrees of freedom, so only models with one or two predictors are considered.

■ Limitations

Limitations on Research

National files such as Medicare, State and local files such as Food Stamps, school enrollments, drivers' licenses, and voter registration lists, and commercial lists were not used. Although some of these files have been difficult to acquire and use in the past (Neugebauer, 1996; Buser et al., 1998) and have not been very useful (Hill and Leslie, 1996; Wurdeman, 1996; Wurdeman and Pistiner, 1997; Sweet, 1997; Larsen, 1999a), few administrative record files were available for this research. Future efforts could acquire a few additional good files for each site or build on other procedures for producing composite population lists (e.g., Sailer and Weber, 1998a, 1998b).

For the files that were available, potentially important characteristics of records were not known. For individual records, the dates of creation and last modification were generally unknown. Geographically associated information, such as estimated poverty, employment, and urbanicity rates, were available at the county level, but not at the block level. Census operation variables like mail response rates, and characteristics from the U.S. Post Office and its Delivery Sequence File (DSF) like percent multi-unit, long-term vacant, and seasonal vacant housing units, were not available at small areas of geography.

For the files and information that were available, the CIFU extract used for this research did not contain as much detail as possible about the differences among fields recorded for individuals in administrative records. Although the sources used for main variables (e.g., first name, last name, age, race) in the composite list are recorded, it is not possible to tell from the research file used which administrative sources contained information and whether or not the versions of an individual's information were identical. Future research files containing the versions of fields for each administrative source would provide more opportunity for analysis (White and Rust, 1997, pp. 65-66; Larsen, 1999a).

In the creation of the composite data base, one address was selected for each person even if multiple addresses appeared in various administrative records. If addresses are lost during file processing, then match rates between census data and administrative records will be hurt, and the match rate is very important to secondary sources, such as administrative records, being useful for coverage improvement (Bye, 1997, pp. 57-59). The low match rate between census and administrative records in 1995 (White and Rust, 1997, p. 74) made it difficult to predict which nonmatched administrative records corresponded to Census Day residents. In 1998, the match rates are higher in part because procedures were used to remove more difficult cases (Owens, 1999). Future research should study the impact of using alternate address information in matching procedures.

Limitations on Ability to Generalize

Although the results are generally positive, they are confounded with site. An experimental design that used at least two sites of each type might provide more believable results. Ideally, several sites of each type, chosen as part of a careful experimental design, would be studied simultaneously.

A related limitation is that sample sizes in the sites are not large. It would be interesting to know how residency rates change with household size, number of administrative records matched to a household, and similar questions, but larger sites are needed to assess these questions.

Further, Census 2000 methodology will differ from that in the Dress Rehearsal, especially in Sacramento where NRFU sampling was conducted.

■ Results

Some results are reported below for individuals in the three sites separately and for the counties of South Carolina. Some comparisons across sites are presented in the Comparisons Across Sites section. More extensive results are contained in Larsen (1999b).

Sacramento, California

In Sacramento, there were 891 residents and 470 nonresidents in the CIFU sample. The groups with the highest rate of residency are defined by having only nonmatched administrative records in the household (rate of 0.839 on 248 records), not having gender recorded (0.828 on 29), being on HUD lists (0.771 on 48), being Asian or Pacific Islander (0.753 on 166), being Black (0.734 on 192), or being another race (0.730 on 111). None of these groups by itself covers more than 20 percent of the file or one-quarter of the residents. The IRS file is the largest and has slightly more residents than on average (0.667 on 1279).

The best combinations of two groups joined together usually include the group that had only administrative record nonmatches in the household with another group that had a higher percentage of residents than overall. Combinations of three groups with the highest percentage of residents generally include the group defined by having only administrative record nonmatches in the household in combination with two other groups, which often include two out of the groups on HUD, on IHS, or with American Indian race. As with two groups, the highest residency rates (max=0.828) include small additions to the group having only administrative record nonmatches in the household.

Logistic regressions were fit, using automatic variable selection to two versions of the data. First, factor variables with indicators for all levels of race, age categories, and Hispanic origin (Hispanic, not Hispanic, not specified) were selected. Second, separate indicators for all levels were entered separately. Then, interactions were considered. The first model selected had effectively 17 variables, whereas the method with separate indicators had 13 variables. When interactions were considered, the two methods produced models each with 26 variables. Table 1 displays the relation of respondent groups based on the predicted probability of being a resident and actual resident status for the main effects only and interaction models using factor variables. Results using individual indicator variables were similar.

The logistic regression equations perform slightly better than the simple groups, but they use much more complicated equations and criteria. Even with the increased complexity, it is not possible to identify most of the administrative records with little error.

In general, the percentage of residents is lower in the cross-validated samples than in the previous column. The validation procedure of applying estimated equa-

Table 1: Number of residents and nonresidents and percentage of residents in data set and in cross-validation in Sacramento, California, by intervals of predicted probability from logistic regression models.

Predicted Probabilities	Residents	Nonresidents	%Resident	%Resident in CV
		Model with Main Effects		
Above 0.9	54	4	0.931	0.876
Above 0.8	240	48	0.833	0.833
Above 0.7	444	100	0.816	0.797
Above 0.6	667	223	0.749	0.723
Above 0.5	792	341	0.699	0.682
		Model with Interactions		
Above 0.9	102	3	0.971	0.812
Above 0.8	259	41	0.863	0.812
Above 0.7	451	100	0.819	0.804
Above 0.6	705	225	0.758	0.718
Above 0.5	781	301	0.722	0.698
Overall	891	470	0.655	0.655

tions to a new site, not just to a test sample from the same site, was not possible.

Columbia, South Carolina

In South Carolina, there were 1,022 residents and 374 nonresidents. The best subgroups are defined by having only nonmatched administrative records in the household (residency rate of 0.901 on 434 records), being of another race (0.893 on 28), being of unknown age (0.795 on 44), being fifty or older (0.790 on 276), and having no matches in the household (0.787 on 1026). The first and last groups mentioned above contain 38.3 percent and 79.0 percent of the residents, respectively. IRS contains most of the records (1,333) and includes slightly more residents than nonresidents (rate 0.746). Race as "other" in South Carolina means, essentially, not black and not white.

As in California, the best two- and three- group combinations usually involve the group defined by having only nonmatched administrative records in the household. There is great variability in the size and composition of resulting groups, but no large improvement is seen.

The number and percentage of residents in groups produced through logistic regressions are similar to the number and percentage produced through the best simple groups. Despite the increase in complexity, not much is gained in ability to identify residents.

Menominee, Wisconsin

Menominee had only 57 residents and 116 nonresidents in the CIFU sample. The best groups are defined by everyone in the household being on multiple sources (residency rate 0.688 on 16 records), being on IRS lists (0.667 on 54), being on multiple sources (0.627 on 51), being on Selective Service lists (0.556 on 9), and having one census nonmatch in the household (0.444 on 54). Race as "other" in Menominee effectively means not American Indian.

In Menominee, Wisconsin, it is possible to identify almost two-thirds of the residents in combinations of two groups that consist of three-fifths to two-thirds residents. There are several different groups involved in these pairs of groups. Most consistently present are IRS, records or households from multiple sources, and Selective Service. No third group can be added to increase the number of residents without greatly decreasing the rate of residency.

The logistic regression model selected for Wisconsin has only one predictor: membership on the IRS list. The sample size at this site prohibits much further discussion.

Comparisons Across Sites

In all three sites, groups defined by being age 0-19, age 50 or more, on IRS, having no matches in the household, and having one census nonmatch in the household had more residents than on average. In Sacramento and South Carolina, females, those on single lists, and those with only nonmatched administrative records in the household, whereas in Wisconsin, males, those on multiple lists, and those with some census records in the household, had more residents than average.

Race categories and whether the household had records from only single, only multiple, or both single and multiple sources were inconsistent. Non-Hispanics were more likely to be residents in Sacramento. However, the comparison could not be made in the other sites. HUD TRACS produced more residents than nonresidents in Sacramento and South Carolina. Surprisingly, IHS was not a good screener for residency in Menominee as it had been in 1995 and 1996 (Larsen, 1999a). Selective Service never produced more residents than usual. In Sacramento and South Carolina, the best two-group combinations included as one group the records from households with only nonmatched administrative records.

Household size and the number of administrative records linked to a housing unit produced inconsistent results across and within sites. For most of the sites, the number of households with many people was small causing the proportions of residents in these groups to be quite variable.

County-Level Covariates

Residency rates among nonmatched administrative records across counties within the South Carolina site range from 57 percent to 98 percent. Richland County is an outlier with the rate of 57 percent and in terms of many other variables recorded on the Planning Data Base (Bruce and Robinson, 1999a, 1999b), which was produced by the Bureau of the Census's Population Division and contains about eighty variables from census operations, estimation programs, and the U.S. Post Office Delivery Sequence File (DSF) on counties in the United States. Richland has the largest number and percent of group quarters, the most movers, the highest percent of Asian and Pacific Islanders, multi-unit dwellings, and most housing units among the eleven South Carolinian counties studied here.

With Richland County removed, the simple linear regressions of residency rate on various predictors are not significant at the 0.05 level, even without adjusting for the number of variables that were tried. The closest are percent long-term vacant (range 0-5 percent, correlation -0.60, slope -0.04, P-value 0.064) and ratio in 1990 of population size to number of housing units (range 2.2 to 2.9, correlation 0.59, slope 0.43, P-value 0.067). Plots of the relationship between rates and the best predictors do not show obvious alternatives to a linear pattern. Some of the other variables have less of a linear relationship to the rate of residency.

Pairs of predictors were used in a two-predictor linear regression of county residency rate. Twelve pairs of variables (out of a possible 2,145 pairs) have values of R-squared above 0.5. The highest two values are 0.697 (predictors percentage poverty and a hard-to-count score from 1990) and 0.620 (predictors percentage renters and percentage population change from 1980 to 1990). A bivariate regression can be significant, but the separate univariate regressions insignificant, because one variable suppresses the influence of the other. Here, due to the extensive searching over pairs of variables and the small number of counties, the results can only be used to suggest that it might be interesting to conduct this analysis on a large sample of counties.

■ Conclusions and Recommendations

Census 2000 Dress Rehearsal Conclusions

It was possible in the three sites to find residents among the nonmatching administrative records at rates higher than in the overall file. Many residents could be found in a few groups, and logistic regressions could be fit and used to increase the chance of finding residents. However, all methods produced groups that included many nonresidents and that excluded some residents. Smaller groups contained higher percentages of residents among their records. The complex methods (e.g., logistic regression with interactions) did better on the data from the site, but did not perform quite as well in cross-validation. It was not possible to test results on new, similar sites.

Comparisons to the 1995 and 1996 Tests

It is difficult to compare the 1998 Dress Rehearsal to the 1995 Census Test (Wurdeman, 1996; Hill and Leslie, 1996) and to the 1996 Community Census Test (Sweet, 1997; Wurdeman and Pistiner, 1997) because the sites, operations, administrative sources, and matching rules were different. Other difficulties in 1995 are described in White and Rust (1997, pp. 66-67).

As in 1998, it was possible in 1995 and 1996 to identify small groups with higher percentages of residents than on average, but large groups tended to contain a lot of nonresidents (Larsen 1999a). IRS files and records for people aged 0-20 had more residents than on average in all three studies. Most of the limitations in 1995 and 1996 are the same as those here.

Census 2000 Administrative Records Research

Census 2000 will include a traditional head count and an Administrative Records Census (ARC) experiment, called AREX 2000 (ARRS 1999), which will test methodology for a census that uses administrative records as the primary source of information. As in past research, the Census Bureau plans to produce information based on a few sites that are chosen to represent a range of enumeration circumstances.

The research here on coverage improvement is relevant to AREX plans because nonmatching administrative records either have to be included or excluded from an administrative records population count. Procedures for assembling the composite file and matching it to the census have significant implications for the quality of the resulting estimates. To the degree that more sites can be used, larger samples collected, and field interviews completed, the better are the chances of learning about the potential for the administrative in counting the population.

In this study, the Planning Data Base with information on counties showed some promise in South Carolina. AREX 2000 should endeavor to collect data on small areas, such as census blocks. Census will be tabulating operations and demographic data at the block level, which could be very useful data for predicting residency status. Additionally, the date a record was created or last modified on the administrative file possibly could be of use. When creating the new data base, the programmers should keep all addresses rather than discarding some when making a composite list, so that efforts can be made to increase the match rate between administrative records and census data (Bye, 1997; White and Rust, 1997; Larsen, 1999a).

■ Administrative Records in the 2010 Decennial Census: Planning Ahead and Experimenting

Cohen, White, and Rust (1999) advocate as top priorities in Census 2000 testing the use of administrative records in an ARC and for coverage improvement. In 2010, it is conceivable that administrative records could be a primary or extremely major source of records for the decennial census.

In the decade after Census 2000, efforts should be made to standardize the acquisition, processing, matching, and followup of administrative records. Additional sources should be matched to a master file to complete the count of the population. An annual list of the population based on administrative sources could be very useful and provide the basis for large-scale studies before the 2010 Decennial Census. Experience should be taken from other agencies, and changes in administrative systems should be suggested to increase the usefulness of administrative records. Edmonston and Schultze (1995, chapter 4), White and Rust (1997), and Steffy and Bradburn (1994, chapter 5) discuss challenges to the use of administrative records to count the population.

■ Acknowledgments

Many thanks to Arona Pistiner, Charlene Leggieri, and the other members of ARRS, Ruth Ann Killion (PRED), Karen Owens (PRED), Ann Vacca (SRD), Margaret Poole (ADMS), NORC, and Datametrics. Partial support was provided by "Estimation Methods," Contract No. 50-YABC-7-66021, Task Order 46-YABC-8-00004, Bureau of the Census.

■ Footnote

[1] The views expressed in this paper belong to the author and do not necessarily reflect those of the Census Bureau or the University of Chicago.

■ References

ARRS (1999), Administrative Records Census Experiment in 2000 (AREX 2000), draft document.

Bruce, A. and Robinson, J.G. (1999a), Documentation for the Planning Data Base.

Bruce, A. and Robinson, J.G. (1999b), The Planning Database: Description and Examples of Its Targeting Capability.

Bureau of the Census (1999a), Census 2000 Dress Rehearsal (1998), on-line information, www.census.gov:80/dmd/www/dress.html.

Bureau of the Census (1999b), Census 2000 Dress Rehearsal Evaluation Summary, U.S. Department of Commerce, Bureau of the Census.

Buser, P; Huang, E.T.; Kim, J.; and Marquis, K. (1998), 1996 Community Census Administrative Records File Evaluation, 1996 Community Census Results Memorandum Series No. 20, U.S. Depart-

ment of Commerce, Bureau of the Census.

Bye, B.V. (1997), Administrative Record Census for 2010 Design Proposal—Final Report, Westat, Inc., submitted to Department of Commerce, Bureau of the Census.

Cohen, M.L.; White, A.A.; and Rust, K.F., eds. (1999), Measuring a Changing Nation.

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

Hill, J. and Leslie, T. (1996), 1995 Coverage Study Results, 1995 Census Test Results Memorandum No. 38, U.S. Department of Commerce, Bureau of the Census.

Larsen, M.D. (1999a), Predicting the Residency Status for Administrative Records That Do Not Match Census Records, Administrative Records Research Memorandum Series No. 20, U.S. Department of Commerce, Bureau of the Census.

Larsen, M.D. (1999b), Predicting Census Residency for Nonmatching Administrative Records, Administrative Records Research Memorandum Series, U.S. Department of Commerce, Bureau of the Census, draft.

Neugebauer, S. (1996), Administrative Records File Acquisition History for the 1995 Census Test, 1995 Census Test Results Memorandum No. 24, U.S. Department of Commerce, Bureau of the Census.

Neugebauer, S.; Perkins, R.C.; and Whitford, D.C. (1996), First Stage Evaluation of the 1995 Census Test Administrative Records Data Base, 1995 Census Test Results Memorandum No. 41, U.S. Department of Commerce, Bureau of the Census.

Owens, K. (1999), Census 2000 Coverage Improvement Followup Operation Matching Specifications, unpublished memorandum.

Prewitt, K.; Shapiro, R.J.; and Barron, W.G. (1999), Evaluation of the Standards for Success, Census 2000 Dress Rehearsal Report Card, U.S. Department of Commerce, Bureau of the Census.

Sailer, P. and Weber, M. (1998a), Household and Individual Data from Tax Returns, U.S. Department of Treasury, Internal Revenue Service, Statistics of Income Research Paper (4/21/1998).

Sailer, P. and Weber, M. (1998b), The IRS Population Count: An Update, U.S. Department of Treasury, Internal Revenue Service, Statistics of Income Research Paper (11/17/1998).

Steffey, D.L. and Bradburn, N.M., eds. (1994), Counting People in the Information Age, National Academy Press, Washington, DC.

Sweet, E. (1997), Evaluation of Using Persons from Administrative Records in the 1996 Community Census, 1996 Community Census Test Results Memorandum Series No. 8, U.S. Department of Commerce, Bureau of the Census.

Thompson, J.H. (1999), Response to the Census 2000 Dress Rehearsal Report Card, Evaluation of the Standards for Success, U.S. Department of Commerce, Bureau of the Census.

White, A.A. and Rust, K.F., eds. (1997), Preparing for the 2000 Census—Interim Report II, National Academy Press, Washington, DC.

Wurdeman, K. (1996), 1995 Coverage Study—Phase II Results, 1995 Census Test Results Memorandum Series No. 52, U.S. Department of Commerce, Bureau of the Census.

Wurdeman, K. and Pistiner, A. (1997), 1995 Administrative Records Evaluation— Phase II, 1995 Census Test Results Memorandum Series No. 54, Revised, U.S. Department of Commerce, Bureau of the Census.