

Executive Summary

Study Overview

To reduce costs many countries use administrative data to assist in censuses or as a replacement to traditional censuses (Farber and Leggieri 2002, Ralphs and Tutton 2011). Currently administrative data are utilized in numerous, critical U.S. Census Bureau programs for population, economic, income and poverty, and health insurance estimates, but administrative data have not yet been extensively used to assist in decennial census operations. The Census Bureau is researching ways in which to use administrative data in decennial census operations to reduce costs.¹ This study, building and expanding on previous research that utilized Census 2000 results, provides a foundation for decennial census operational research on administrative records by assessing the quality and coverage of administrative data relative to the 2010 Census.

In the United States, decennial censuses determine apportionment of state representation to Congress, are used in state redistricting, and are used to distribute billions of federal dollars (Reamer 2010). While households are required by law to participate in the decennial census, there are many households that do not respond to initial contact attempts. This requires the Census Bureau to send enumerators door to door to collect data from non-responding households in decennial census operations called Nonresponse Followup Operations.² This effort is expensive for the Census Bureau and was estimated to cost around 1.4 billion dollars in Census 2000 of a total census budget of six billion dollars (Farber and Leggieri 2002, Walker et al. 2012). The estimated cost of these operations in the 2010 Census was about two billion dollars (Walker et al. 2012). Administrative records may be able to assist with expensive operations such as Nonresponse Followup Operations, which would save the government and taxpayers a substantial amount of money.

Census Bureau staff conducted research on the use of federal administrative data utilizing Census 2000 results. The Statistical Administrative Records System (StARS) was developed from select federal data sources in 1999. Decennial census research using these data included address and person count comparisons relative to Census 2000 (Farber and Leggieri 2002). StARS 1999 was also utilized in a field test that simulated a census in several counties during Census 2000 (Berning 2003, Bye and Judson 2004).

The 2010 Census Match Study builds on this research by evaluating the federal data sources used in StARS, additional federal data sources, and commercial data. This report is also distinctive from past research in that it matches addresses and persons in administrative records to the 2010 Census to evaluate the quality and coverage of administrative data. The matching is conducted

¹ For the purposes of this report, “administrative data” and “administrative records” are used interchangeably.

² Nonresponse Followup Operations include Nonresponse Followup, Nonresponse Followup Reinterview, Nonresponse Followup Vacant Delete Check, and Nonresponse Followup Residual. For more information, see Walker et al. (2012).

using unique address and person identifiers called master address file identification numbers and protected identification keys assigned by the Person Identification Validation System to addresses and persons in the 2010 Census and administrative records. Using count and match ratios, this study evaluates the administrative data and the 2010 Census at different levels of geography and by factors such as Hispanic origin, race, and mode of data collection. This report also evaluates the quality and coverage of Hispanic origin, race, sex, and age response data in administrative records relative to the 2010 Census.

Results Overview

Addresses

There were 131.7 million addresses in the 2010 Census and 151.3 million addresses in administrative records. Of the 2010 Census addresses, administrative records matched to 122.0 million or 92.6 percent; 29.3 million administrative records addresses were not found in the 2010 Census; and 9.7 million addresses were in the 2010 Census, but not in administrative records.

Definitional differences between addresses in the 2010 Census and administrative records contributed to the address non-matches. For instance, there were Post Office Box addresses in administrative data but none in the 2010 Census. The 2010 Census also contained physical descriptions for addresses such as “yellow house near fork in the road” that cannot be matched to administrative records. Additionally, administrative records contained non-residential addresses and may have contained new construction that was not recorded in the 2010 Census.

Persons

The person match ratios were lower than the match ratios for addresses. This is in part because all addresses in the 2010 Census had master address file identification numbers, thus all 2010 Census addresses had the potential to be matched to administrative records addresses with master address file identification numbers. However, in the 2010 Census, not all persons received a protected identification key, reducing the number of persons in the 2010 Census that had the potential to match to administrative records. Protected identification keys were assigned through probabilistic matching to records using name, address, and date of birth information.

There were 308.7 million persons in the 2010 Census, and 279.2 million were assigned a protected identification key. There were 312.2 million unique persons in administrative records that were assigned a protected identification key and were alive on Census Day, April 1, 2010. Administrative records matched to the vast majority of persons in the 2010 Census that received a protected identification key, 273.6 million or 98.0 percent. The percentage of the entire 2010 Census universe, including records lacking protected identification keys, with matching administrative records was lower at 88.6 percent.

There were 29.6 million 2010 Census persons that did not receive a protected identification key. There were 48.8 million administrative records that were assigned a protected identification key, but did not match to the 2010 Census. Future research will study the potential overlap between these universes.

There were 5.5 million 2010 Census persons with protected identification keys that were not found in administrative records data, and most of them were under the age of 17. There were several reasons why administrative data did not cover children as well as other age groups, including timing issues with tax data. Tax return data from the previous tax year failed to include babies born after January 2010, however these children would likely be reported in the 2010 Census, resulting in a lower match between administrative records and the 2010 Census for babies.

Person-Address Pairs

The match ratios for person-address pairs (i.e. a person at an address) were lower relative to the address results and person results, in part because the person-address pair data incorporate both address and person matching issues, including the presence of multiple addresses for persons in administrative records. Of the 312.2 million persons in administrative records that had a protected identification key, 301.5 million had a master address file identification number and 10.7 million did not have a master address file identification number. There were 216.2 million person-address pairs in the 2010 Census that matched to administrative records. Of the 308.7 million persons-address pairs in the 2010 Census, 70.0 percent matched to administrative records person-address pairs. Of the 279.2 million person-address pairs in the 2010 Census that had a protected identification key, 77.4 percent matched to administrative records person-address pairs.

After the best address model was applied to persons in administrative records with multiple addresses in administrative records, there were 203.2 million person-address pairs in the 2010 Census that matched to administrative records. Of the 308.7 million persons in the 2010 Census, 65.8 percent matched to administrative records person-address pairs. Of the 279.2 million person-address pairs in the 2010 Census that had a protected identification, 72.8 percent matched to administrative records person-address pairs. There were 98.6 million administrative records person-address pairs that did not match to the 2010 Census. There were 76.0 million person-address pairs that were in the 2010 Census which did not match to person-address pairs in administrative records.

Demographic Quality and Coverage

The quality of Hispanic origin response data from federal and commercial files, as defined by response match ratios between the 2010 Census and administrative data, ranged from 29.4 percent to 93.1 percent. Overall, federal data sources tended to have higher quality race data for each race group relative to the commercial data. The quality of race data varied by race group.

The White alone, Black alone, and Asian alone populations tended to have higher quality race data in administrative records compared to Two or More Races, Native Hawaiian or Other Pacific Islander alone, American Indian or Alaska Native alone, and Some Other Race alone populations.

Federal and commercial files had high quality data for age and sex responses. Across federal and commercial files that had date of birth information, the age match ratio ranged from 79.0 percent to 98.5 percent. The sex match ratios ranged from 94.7 percent to 100.0 percent.

The demographic coverage analysis evaluated whether administrative data provided a demographic response to Hispanic origin, race, age, and sex groups in the 2010 Census regardless of the quality of the response. There was a Hispanic origin response present in administrative data for 92.2 percent of non-Hispanic respondents and 78.9 percent of Hispanics in the 2010 Census. The race response coverage in administrative records ranged from 46.1 percent for the Some Other Race alone population to 81.0 percent for the White alone population. Coverage by age group ranged from 84.9 percent to 94.3 percent with older age groups achieving higher coverage relative to younger age groups. Coverage for sex was 90.1 percent, where females had slightly higher coverage (90.8 percent) relative to males (89.3 percent).

Research Implications

1. **Administrative records can enhance, but not replace the decennial census.** While the quality and coverage of administrative records relative to the 2010 Census suggests that administrative records can be utilized in decennial census operations, the quality is not high enough and the coverage is not expansive enough to replace a traditional census.
2. **Use of administrative records in Nonresponse Followup can reduce costs.** Administrative records cover a substantial number of Nonresponse Followup addresses and persons, and nearly half of person-address pairs. Of the 23.6 million addresses that responded in Nonresponse Followup in the 2010 Census, administrative records matched to 21.0 million or 89.2 percent.³ Administrative records also matched to a substantial number of persons that were in Nonresponse Followup in the 2010 Census. Of the 60.4 million persons in Nonresponse Followup in the 2010 Census, 48.0 million or 79.5 percent were in administrative records. Administrative records matched to a lower number and proportion of person-address pairs in Nonresponse Followup compared to addresses and persons. Of the 60.4 million 2010 person-address pairs in Nonresponse Followup, there were 28.7 million or 47.5 percent that matched to administrative records.

³ There are 47.2 million housing units in Nonresponse Followup according to the “2010 Census Nonresponse Followup Operations Assessment” (see Walker et al. (2012)). This number is much higher relative to the housing units in this report for several reasons. For instance, the number of Nonresponse Followup housing units in Walker et al. (2012) include vacant, deletes, and unresolved households, whereas the Nonresponse Followup housing units in this report are all occupied.

Research and improvements in record linkage, refinements of the best address model, and acquiring data that cover those most likely to be in Nonresponse Followup may enhance the person-address match between the 2010 Census and administrative records.

3. **Administrative records can assist in determining housing unit and occupancy status.** Administrative records can assist to verify whether a housing unit is a valid livable housing unit and whether it is occupied. Occupancy status results demonstrate the value of administrative records for these purposes. Of the 116.7 million occupied housing units in the 2010 Census, administrative records indicated that 96.1 million or 82.3 percent were occupied. The 2010 Census designated 15.0 million housing units as vacant, of which administrative records found that 11.4 million or 76.1 percent were not occupied. Of the 4.9 million housing units designated as deletes in the 2010 Census, administrative records indicated that 4.2 million or 85.4 percent were not occupied.⁴
4. **Administrative records can inform household population count assignment.** Administrative records had the same population count for the majority of 2010 Census housing units that matched to administrative records. Of the 116.7 million 2010 Census occupied housing units, 96.1 million matched to administrative records. Of these, 55.5 million or 57.7 percent of housing units had the same population count. When administrative records and the 2010 Census did not have the same population count, the count differed by one person for 63.7 percent of the housing units. Further research should be conducted on this universe.
5. **Acquiring additional federal, state, and commercial data can improve address, person, and demographic characteristic coverage.** Administrative data do not cover children as well as they cover adults. Also, the quality of race and Hispanic origin response data from federal and commercial sources varies considerably by race and Hispanic origin group. The Census Bureau should partner with federal agencies, state agencies, community groups, and other organizations to obtain data that contain information on children living in households, and additional race and Hispanic origin response data should be acquired, particularly for groups where the quality of race or Hispanic origin response data is low in administrative records. Obtaining data for the following groups should be a priority: Two or More Races, Native Hawaiian or Other Pacific Islander, and American Indian or Alaska Native.
6. **Administrative records can inform race and Hispanic origin determination.** For some race and Hispanic origin groups, the quality of administrative records response data was high. For instance, the White alone, Black alone, and Asian alone populations had

⁴ Deletes refer to housing units designated for deletion from the address list. Housing units may be identified as deletes for a number of reasons including being demolished, uninhabitable, or nonresidential. Counts of 2010 Census addresses designated as deletes may vary across 2010 Census Program for Evaluations and Experiments reports as a result of different data sets being used for analysis.

relatively high quality race response data in administrative records compared to other race groups. The quality of administrative records files ranged from 94.7 percent to 99.1 percent for the White alone population. The quality of federal data for the Black alone population ranged from 87.4 percent to 98.3 percent. The range was considerably lower for commercial data. For the Asian alone population, the quality of both federal and commercial data ranged from 58.0 percent to 94.1 percent. Data could also be used for other race groups from administrative records, but the quality was generally lower. Research should be conducted on how administrative records can assist with race and Hispanic origin determination for censuses and surveys.

7. **Administrative records can assist age and sex determination.** The quality of age and sex response data in administrative records is high. For sex, the quality of administrative data ranged from 94.7 percent to 100.0 percent across administrative records files. For age, in data sources that contained date of birth, the quality of administrative records ranged from 79.0 percent to 98.5 percent. Research should be conducted on how administrative data can assist with age and sex determination for censuses and surveys.
8. **Conduct additional record linkage research with the aim of improving match results for unvalidated person records.** Many improvements were made to the Person Identification Validation System to enhance the assignment of protected identification keys and master address file identification numbers to administrative records data. Continued record linkage research on the Person Identification Validation System should be conducted to further enhance the assignment of protected identification keys and master address file identification numbers to persons and addresses, potentially increasing the universe of persons and addresses that can be matched and unduplicated between censuses and surveys and administrative records. For instance, of the 308.7 million persons in the 2010 Census, 29.6 million did not receive a protected identification key. Of these, 10.3 million could not be sent through Person Identification Validation System processing because they lacked name and date of birth, and 19.3 million went through Person Identification Validation System processing but failed to receive a protected identification key. Additional research should be conducted on how to minimize this latter universe.
9. **Conduct record linkage research to improve match results for records with incomplete name and date of birth data.** Commercial data sources often lack complete name and date of birth information. Research to unduplicate these records that failed the Person Identification Validation System, and assess the quality of the data is needed. Research on how to use records that lack personally identifiable information is needed, moving the matching approach beyond validation using the Social Security Administration Numerical Identification File.

10. **Conduct record linkage research that improves person record unduplication.**
Current record linkage techniques must determine whether two people that look similar are indeed the same person or if they are two different people. Refinements on record linkage techniques will help to more accurately unduplicate person records.
11. **Develop partnerships with federal and state agencies to better understand administrative records and enhance record linkage research.** Partnering with federal and state agencies will facilitate knowledge sharing on the availability of data that could enhance record linkage processes. This knowledge sharing will also benefit administrative records research. For instance, a better understanding of how data were collected could assist in the validation and unduplication process and improve understanding of resulting linkages.
12. **Assess whether an administrative records composite improves missing data assignment.** Building an administrative records composite involves unduplicating records, assigning persons at multiple addresses to one address, and assigning one characteristic to people that have different characteristics across source files. Research should assess the quality of missing data assignment using a composite compared to using all available administrative data.
13. **Analyze linked survey data, especially the American Community Survey, to explore characteristics associated with data coverage and consistency.** Evaluating administrative records relative to the 2010 Census provided important information, at different levels of geography and by certain characteristics, about the quality and coverage of administrative data. Other evaluations using survey data such as the American Community Survey can provide additional insights because the American Community Survey has many additional characteristics that can be analyzed.