TOWARDS A U.S. POPULATION DATABASE FROM ADMINISTRATIVE RECORDS

Kent Marquis, Signe Wetrogan and Henry Palacios, U.S. Census Bureau*
Kent Marquis, U.S. Census Bureau, Washington, DC 20233-9150

Key Words: 2000 Census, Coverage, Record Linkage

1.0 INTRODUCTION

In visions of future government statistical systems, administrative records will be used widely. Demand for information will increase beyond the ability of surveys and censuses to meet it. Administrative records will become easier to get and process, enabling more and cheaper estimates with less burden on respondents.

But administrative records are a by product of systems that we currently neither fully understand nor control, so many concerns arise about our ability to use them and the quality of estimates based on them.

Last January, the authors conducted a review (Marquis, Palacios, and Wetrogan 1996) and concluded that many of the Census Bureau's programs would benefit by creating a national database of people linked to families, addresses, and geographic areas. This paper raises some orienting issues concerning a national population database, provides some early results related to a few of those issues, and discusses next steps for moving ahead with database building.

2.0 BACKGROUND

2.1 Forces Promoting Change

The Federal Statistical System is experiencing a number of external forces that point it toward a greater use of administrative records. Spending on Federal activities is decreasing. Governments, themselves, are developing more administrative databases. And the Paperwork Reduction Act encourages us to cut the burden on respondents.

We have information age tools such as hardware, software and networks to facilitate using administrative records.

National political leaders are working out a "devolution" of responsibility from the federal to state and local levels in such areas as social welfare programs. Devolution increases the need for small area data that administrative records might supply.

The general public expects that agencies will use and reuse available data rather than collecting it again.

Next, we will outline some of the advantages and problems of using administrative records instead of the primary survey and census data that we usually use.

2.2 Advantages of Administrative Records

Efficiency is one advantage since using administrative records can be less expensive than collecting primary data. Response burden on citizens and organizations can be reduced.

Since administrative information is about all population units, we can expect more precise small-area estimates. Administrative records databases mean that potentially more information is available to statistical agencies, unconstrained by the length of questionnaires.

Rates of missing data can be lower on administrative records. This means fewer nonresponse adjustments, fewer imputations and potentially less nonobservation bias.

Administrative records can diminish measurement problems. There are no more interviewer effects. One can get better information about sensitive topics, can overcome deliberate lying and bypass refusals to be counted. And we can avoid human cognition effects such as forgetting, poor comprehension, and laziness.

2.3 Problems with Administrative Records

But, statisticians see an almost overwhelming collection of problems with administrative records. The problems fall into several categories:

The first is access: Administrative records can be difficult or impossible to get. Costs can be surprisingly high for secondary data: to locate, to purchase, to match and review, to fix errors and to produce estimates.

In the U.S., the population coverage of most record systems is incomplete. It is unclear what happens to coverage when record systems are combined.

The quality of measurement, especially for the information that is not critical to the primary administrative use, can be poor.

Seldom will the administrative record system be measuring the exact concept needed by the sec-

*This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

ondary user. For example, Social Security records code 4 race categories. The 1990 Census used more than a dozen.

Timeliness: Data from administrative records systems often are not current or current information is not yet complete.

Matching: Records must be linked with other records to estimate relationships among variables. Missing and faulty data can cause match errors and distort estimates. Statistical matching often requires human intervention to resolve ambiguous cases. Archiving and storage of masses of administrative information becomes an issue.

The privacy, confidentiality and disclosure issues for administrative records are basically the same as for primary data. One exception is informed consent, because people who provide the data in administrative records aren't always told about statistical uses. In addition, protecting against disclosure of individual information is becoming more challenging.

Administrative records databases are used successfully in other countries such as Canada (Standish and Ravindra, 1996) and Finland (Myrskyla, 1991 and Lankinen, 1995). So it is appropriate for the U.S. Census Bureau to begin a program of research and development in this area. In the next section we look at some of the beginning issues concerning the quality of data on administrative files.

3.0 RESULTS

3.1 Are Addresses on Tax Forms Residential?

First we ask about the quality of addresses on the individual federal tax files. Sater (1995) matched the national tax files to households interviewed in the Current Population Survey (CPS) at their residences. If the 2 records matched on social security number, he compared the house number and street names in both records, declaring a match if the 2 fields matched exactly or the street names matched and a human judged the house numbers to be the same. For this subset, 86% of the tax form addresses were essentially the same as the CPS residence address.

For most Census Bureau uses, addresses need to be assigned a code that locates them in a small geographic area, for example, a particular side of a city block. Wetrogan et al. (1996) attempted to assign codes to addresses in a 1% sample of a recent national tax file. For states with a large proportion of city style addresses, the computer could assign codes to about 90% of the addresses. Results for other states were not so encouraging.

3.2 Quality of Addresses on Other Federal Files

We can examine the quality of addresses on other federal files. In a recent study in one city (Moore, Marquis and Bogen, 1996), we drew samples from lists of people receiving selected government transfer payments. Interviewers contacted each address and listed the people living there. We then determined whether the person sampled from the administrative records was residing at the address.

Record Source	Sample Person Found at Address
AFDC	73%
Food Stamps	73%
Supplementary Security Income	75%
Unemployment Compensation	79%

Table 1. About 3/4 of the Sampled People Were Found at the Address in Administrative Records.

Table 1 indicates that about one quarter of the sampled people were not residing at the address listed in the administrative records.

These studies evaluated addresses on single administrative records files. Results might be better if we combined information from several administrative record sources into a database. We will look at this next.

3.3 Test Census Address Matching Results

In 1995, we conducted Test Censuses of population in 3 areas of the U.S. As described by Neugebauer, Perkins and Whitford (1996), we also built an administrative records database for each test area.

We used a wide range of administrative files including food stamps, tax returns, drivers' licences, voter registrations, public housing, school enrollment and parolees.

We matched on several combinations of variables such as last name, date of birth, and social security number. One version of the street address was retained. This was a first attempt to build the administrative records databases and gain experience with a variety of files. The plan is to evaluate

this beginning attempt and then to rebuild the databases, from scratch, using the knowledge gained in Stage 1.

Next, we matched the databases to the Test Census address files on the complete street address to see how much direct overlap there was.

The match rates in Table 2 show quite a bit of intersite variation when we matched on the full address. However, these sites were selected, in part, because they represent very different kinds of address problems. The 6 Louisiana parishes are rural and addresses are often rural routes or post office boxes. The Census list is often in terms of location descriptions such as "White frame house, ½ mile down dirt road." We would not expect high match rates under these circumstances. Oakland is more typical of urban America. Both records and Census lists are in terms of house numbers and street names. There, about 2/3 of the administrative records full addresses matched exactly to the Census address file. One characteristic of the Paterson test site was the large number of multiunit structures. Subunits are not always identified by administrative records address information. Census lists, however, include explicit subunit identifiers. Thus, low address match rates were expected (and obtained) in Paterson when matching on the full address (30% match).

Test Site	Percent Administrative Records Addresses Matched to Census Using:	
	Full Address	Basic Street Address
Paterson NJ	30%	70%
Oakland CA	66	78
6 Louisiana Parishes	24	Not Analyzed

Table 2. Administrative Address Match Rates to Census Address Lists in the 3 Test Sites, for 2 kinds of match requirements.

To look at the extent of the unit designator problems, we rematched the databases to the Censuses in Oakland and Paterson, on just the basic street address, which does not require the unit designators to agree. More matches were made, especially in Paterson where the match rate more than doubled from 30 to 70 percent.

Taken together, despite using very different methods, these 4 studies all suggest that there will be some problems and discrepancies with addresses obtained from administrative records. Extra steps may be needed to obtain more complete address information (including unit designations) and also to validate addresses before using them to classify a person's geographic status for estimation purposes.

3.4 Towards A Population Database

Now we'll switch from addresses to people. The administrative records databases may be viewed with the person rather than the address as the basic unit. To evaluate the person databases, they were matched to the Census in each of the 3 test sites, using name, date of birth, and other variables. The detailed results are in the Appendix. Here (Table 3) we show a composite set of preliminary results.

In Census?	In Administrative Records Database?		
	Yes	No	**Total**
Yes	2	1	3
No	4		
Total	6		

Table 3. Person Match Results, Composite for 3 Test Sites

Table 3 contains the relative frequencies of people in each of the match classification cells. Matched people are in the upper left cell. In each site, for every 2 matched people, there were about 4 people in the database who could not be found in the Census and there was approximately 1 person in the Census who could not be found in the administrative records database. Looking at the marginals, there were about twice as many people in each database as in the corresponding Census.

4.0 DISCUSSION

To examine the implications of these results, consider 3 potential uses of the population databases:

1. Adding people to the Census who were missed by the usual counting procedures
2. Using administrative records as a source of information about households who did not respond to the Census questionnaire.
3. Using the administrative records database to

estimate population sizes.

The first 2 uses are part of the current plans for the 2000 Population Census.

Consider adding missed people to the Census: At the end of a reinterview with a sample, one possibility is to ask the household about any people listed in the administrative records but not on the Census or reinterview. Verified, additional people would be added to the household and used in subsequent estimates.

The benefits of this procedure are to increase the quality of the Census counts and to reduce what we suspect is an undercount by traditional procedures.

Indeed, in a special study (Hill and Leslie, 1996), the Census actually tried to locate a sample of the people who appeared only in the administrative records database. In Paterson about 32% of the sampled people were found and in Oakland, 18%. While these are probably overestimates of true Census misses, the potential for adding real people from the administrative records database is still great.

But there are also risks. One is that the differential undercount might actually increase if administrative records add disproportionately more people from majority groups. A second risk is adding too many people to the Census, because of match error, followed by imperfect reconciliation at the household.

Indeed, match error may be the Achilles heel of administrative records estimation and we digress briefly to consider it. Marquis (1978) showed how match errors can distort estimates and adversely affect logical inferences from applications based on just one of the off-diagonal error cells. Here let us consider three of many practical examples (Note that field procedures may prevent some of these literal situations from happening. But we use them anyway to illustrate the general points):

Using the 2 - by - 2 cross classification that results from matching 2 lists, consider the example of a person who is known by two versions of a given name, say Randolph and Sonny, and assume that the name standardizer, part of the preprocessing activities, does not know that Randolph and Sonny are variations of the same name.

Census	Administrative Records Database	
	Yes	No
Yes		Randolph
No	Sonny*	

*Asked about during the reinterview, potentially added.

CASE 1: One scenario (above) is to have one record for Randolph in the Census and one record for Sonny in the administrative records database.

Randolph and Sonny are the same person. But they could not be linked when we matched the Census and the administrative records database. So the household was revisited and asked whether Sonny lived there on Census Day. We hope we had good procedures and an alert respondent, and a careful enumerator and that they figured out that Sonny was a duplicate.

CASE 2: A different scenario is to have two records for the target person in the database that could not be unduplicated, and one correct record in the Census.

Census	Administrative Records Database	
	Yes	No
Yes	Randolph	
No	Sonny*	

*Asked about during the reinterview, potentially added.

When the database is matched to the Census, one of the database records will match and one will not. In this example, Sonny was not found in the Census. So, during the follow-up, the enumerator might ask if Sonny lived there on Census Day. The answer would be "yes" and we risk adding an inappropriate record to the Census unless the procedures work and Sonny is recognized as a duplicate.

CASE 3: A final example to make a point about definitions and data quality. Let's suppose Randolph is in the database once and his address is 101 Main Street. Let's suppose Randolph was in jail on Census Day and that he was not listed on the Census questionnaire as a resident.

Census	Administrative Records Database	
	Yes	No
Yes		
No	Randolph*	

*Asked about during the reinterview, potentially added.

When the follow up interviewer asks if Randolph lived there on Census Day, we hope that the respondent is able to accurately reconstruct and report that Randolph was not there.

To remove the risk of inappropriately adding a person to the Census, we might consider special treatment for people in special circumstances. In this case, if the database had a list of people in prison on Census Day, it could be programmed to ignore records for prisoners that say they lived elsewhere.

The general concluding point is that it is better to prevent the errors from occurring in the first place, than to rely on later procedures to detect and correct them.

A second use of administrative records is as a source of information about households who don't mail back their Census form. The Census sends human enumerators to follow-up each non responding household. This is costly and causes delays in the processing and release of information. For the 2000 Census, research is looking into using administrative records for at least 5 percent of the nonresponding households.

Risks involve adding incorrect information because the database contained too few or too many people at the household or because the information about the people was not correct.

A third use of administrative records databases is to make estimates of population sizes, both nationally and for small areas. The traditional method of determining population size is to conduct a census. But this is expensive and not everyone makes the effort required to be counted. Rates of public cooperation seem to be on a downward trend, while costs of traditional counting continue to increase. It is unclear that there will be an available labor pool large enough to conduct the next Census in the United States using traditional methods.

While the Census Bureau does not propose to use administrative records to conduct the 2000 Census, there are other applications to consider. One is the 2010 Census, others involve updating the estimates of population for small areas between censuses to facilitate equitable revenue sharing and for other purposes.

One risk is that we will overestimate population size because of the net excess of false positive people in the database. Recall from Table 3 that for every 2 matched people, there were 4 people only in the administrative records database and 1 person only in the Census. So there are many more people in the database than in the Census, here, about twice as many. If we were to base the population estimate on the database, the estimate would be twice as large as the count based on the Census.

These evaluations suggest that the administrative records databases can be improved. Current plans are to rebuild the Test Census databases and to do things a little differently.

We expect to reduce false positives by not using files or records that contain missing or faulty match information. We will also make some improvements in the name standardizer.

The new matching tactics are: (1) to use model-based, statistical matching techniques and (2) to unduplicate people records within each address first, then unduplicate people across addresses.

Over the long term, even more difficult steps may be needed. For example, we need to work with data suppliers to assure higher quality data in critical matching fields, we need to acquire lists of people in group quarters and special places to make sure they are represented only in the appropriate geographic areas. We need ways of identifying transitioning people such as newborns, movers, seasonal residents, people who change their names, and deaths. International immigration and emigration data may be needed at the person and family level.

A full set of address change data may be needed annually and it may need to be processed longitudinally. New construction and building conversion information will be needed periodically. If specific undercoverage problems are found, we will need to include additional administrative files in the database to improve coverage.

And when we solve the coverage problems, we can go on to worry about measuring content.

In sum, there are good reasons for the Census Bureau to explore using administrative records in a big way, such as creating a population database. But there are many problems to overcome, such as incomplete and faulty address information. Our Phase 1 attempts to build the population databases for the 1995 Test Census Sites taught us a lot and revealed that we need to be concerned about over-including people as well as omitting them. Both kinds of match errors pose risks for the 3 uses we examined. We discussed next steps to rebuild the databases in Phase 2. We will try some new things and will try to remember that small is beautiful, even on a national scale.

REFERENCES

Hill, Joan and Theresa Leslie (1996), "Coverage Study Results," 1995 Census Test Results Memorandum No. 38, U.S. Census Bureau, Washington DC, March 13, 1996.

Lankinen, Markku (1995), "Measuring and Forecasting Housing Needs at Regional Level in

Finland: Quantitative and Qualitative Aspects. Case of Helsinki," Proceedings. 50th Session, International Statistical Institute, Beijing.

Marquis, Kent (1978), "Inferring Health Interview Response Bias from Imperfect Record Checks," Proceedings of the Survey Research Methods Section, American Statistical Association, 265-270.

Marquis, Kent, Henry Palacios and Signe Wetrogan (1996), "Draft Recommendations for Establishing the Census Bureau's Administrative Records Capability." Report prepared for the Team for Administrative Records Planning. U.S. Census Bureau, Washington DC, Jan. 29, 1996.

Moore, Jeffrey, Kent Marquis and Karen Bogen (1996), "The SIPP Cognitive Research Evaluation Experiment: Basic Results and Documentation," Center for Survey Methods Research, U.S. Census Bureau, Washington DC.

Myrskyla, P. (1991), "Census by Questionnaire, Census by Registers and Administrative Records: The Experience of Finland," J. Official Statistics, 7, 457-474.

Neugebauer, Sharon, R. Colby Perkins and David C. Whitford (1996), First Stage Evaluations of the 1995 Test Administrative Records Database," 1995 Census Test Results Memorandum 41, U.S.Census Bureau, Washington DC, March 14, 1996.

Sater, Douglas (1995), "Differences in Location of Households and Tax Filing Units," Paper presented at the Annual Meetings of the Population Association of America, April 1995, San Francisco

Standish, Linda and Daniela Ravindra (1996), "An Overview of the Longitudinal Administrative Databank: Evaluation of Representativity," Proceedings of the 1996 Annual Research Conference, U.S. Bureau of the Census, forthcoming.

Wetrogan, Signe et al.(1996), "Administrative Records: Panacea or Pandora's Box?," Paper presented at the Annual Meetings of the Population Association of America, May 1996, New Orleans.

Acknowledgment

We would like to acknowledge the courageous people who built the population databases: Billy Stark, Sharron Baucom, Vickie Kee and the administrative records processing staff. They succeeded in spite of government shutdowns, management changes, office moves, a severe automobile accident and other adversities.

APPENDIX

Match Cross-Classification Frequencies (000's) of People in The Administrative Records Database and/or in the Census, by 1995 Test Census Site.

Test Census Site	Match*	In Census Only	In Adm. Records Only	Total Census	Total Adm. Records
Paterson	72	48	138	120	210
Oakland	199	108	531	307	730
LA Parishes	81	38	157	119	238

*The Census and administrative records estimates are from slightly different populations (see Neugebauer, et al, 1996, for details). In cases where the estimated number of matched cases was not exactly equal for each source, we used an average. .