Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Using Administrative Records to Evaluate Survey Data

Mary H. Mulry, Elizabeth M. Nichols, and Jennifer Hunter Childs[1]

Abstract

After the 2010 Census, the U.S. Census Bureau conducted two separate research projects matching survey data to databases. One study matched to the third-party database Accurint, and the other matched to U.S. Postal Service National Change of Address (NCOA) files. In both projects, we evaluated response error in reported move dates by comparing the self-reported move date to records in the database. We encountered similar challenges in the two projects. This paper discusses our experience using "big data" as a comparison source for survey data and our lessons learned for future projects similar to the ones we conducted.

Key Words: recall error; memory, telescoping, move date

1. Introduction

Administrative records provide a source of data to use to evaluate errors in survey responses. Such evaluations have the potential to aid in the design of data collection but also to provide insight for designing estimation methodology that relies on a combination of survey and administrative records data or a transition from survey data to an administrative source. Although administrative records contain copious amounts of data, they are collected for their own purposes and have their own sources of error. Using administrative records to evaluate survey data is not always as straight forward as it may sound. In this paper, we draw on our experience with two studies of error in survey reports of moves that used administrative records to discuss some of the challenges that researchers must address when using administrative records to evaluate survey error (Nichols, Mulry, and Childs *forthcoming in 2017*). These two studies are a small part of the U.S. Census Bureau's large research program aimed at increasing the use of administrative records and third-party data sources in its census and survey products (O'Hara 2014).

For context, the United States has a decentralized statistical system. The U.S. Census Bureau is the largest federal statistical agency, but other federal, state, and local agencies collect and hold data. Some of the federal agencies are statistical in nature, but other agencies have missions that are not statistical by design -- for example, the U.S. Postal Service. In addition, other entities, including commercial or third party entities and universities, also collect and maintain data. Each of these stewards collects data for their own purposes, not the Census Bureau's purpose.

To use data from other organizations, the U.S. Census Bureau enters into agreements with these other entities on how we will use and protect these data. These agreements often take years to develop. There are physical and legal restrictions on using these data once at the U.S. Census Bureau (Johnson, Massey, O'Hara 2015).

The U.S. Census Bureau does use administrative records in the course of producing some of its economic and demographic estimates. However, the U.S. does not have a continuously updated list of people residing within its boundaries. The U.S. decennial census occurs every 10 years and has been based on reports from the people

[1]Mary H. Mulry, Elizabeth M. Nichols, and Jennifer Hunter Childs, U.S. Census Bureau, Washington, DC 20233. This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

themselves, not administrative records. More recent advances in computing technology and electronic storage capabilities has led to an expansion in the U.S. Census Bureau's research into using administrative records and third-party databases with the goal of reducing cost, increasing efficiency, improving quality, generating new products, and reducing respondent burden.

2. Research Focus

Our work on the two studies concerning the evaluative use of administrative records and third-party data started with the research question:

> What is the measurement error in survey reports of dates of residential moves as the length of time between the move and the interview increases?

The answer to this question is important to both the U.S. Census Bureau and survey researchers. The U.S. Decennial census assumes people who did not initially self-respond remember where they lived on census day (April 1 of the census year) when the non-response follow-up interviewers contact them two or three months later. In addition, evaluative operations that occur several months later also make this assumption. Survey researchers have made similar assumptions about respondents' recall of move dates when they use moves as anchors to aid memory of events, particularly in the technique that creates an event history calendar (Belli 1998).

This paper focuses on two methodological projects we did in an attempt to answer the research question. We used administrative records and third party data as the comparison source to a survey report. In a sense, this paper focuses on the feasibility of doing these types of evaluations and what we learned during the implementation of our studies as opposed to presenting the results we found.

The first project was conducted under contract with NORC (Krishnamurty 2012). We used three selected years from the National Longitudinal Survey of Youth (NLSY) as our survey source. The NLSY is an annual longitudinal, in-person survey and asks respondents for the dates of all move between cities since the prior interview. When the interviewer has difficulty finding a sampled respondent for the next wave of interviews, they turn to a third-party database called Accurint a locate-and-research tool owned by Lexis-Nexus to help locate the individual (Mulry, Nichols, Childs 2014). Accurint maintains records for over 400 million identifiers and compiles data from over 10,000 sources. These records associate a person's name with an address and dates at that address.

The second project was conducted at the U.S. Census Bureau. In this project, we sampled from an administrative record source and then conducted interviews of the people in sample. The source was the National Change of Address (NCOA) files from the U.S. Postal Service (USPS; U.S. Postal Service 2015). We selected a sample from NCOA forms files in March & April of 2010, the census year, and conducted a telephone interview with the household in later months, simulating the timing of the different census operations.

That interview asked about other places where each household member lived during the year and the dates of the move. In this project, we also compared dates from the change of address file to the reported move dates
We experienced similar challenges with both projects, and those challenges are relevant to other studies using administrative records as the source of data for an evaluation of survey responses.

When identifying a data source to use in an evaluation, "fitness for use" always presents a challenge. Our two surveys asked for residential move dates, but a date in the Accurint database was not necessarily a move date nor was the NCOA date that requested forwarding of mail. However, we had to consider whether this mattered for our research question. We concluded that in this evaluative study, the definitional challenges remain constant over time, and given that assumption, we still felt confident that our results of whether time since the move affected the memory of that move was still viable. We were not measuring the number of moves a person made by these records, but if we had been, our conclusion may have been different.

3. Matching

One of the first things we needed to do for both projects was to link or match the administrative or third-party data to the survey data. This proved to be one of the most challenging aspects of both studies.

First, we focus on the Accurint/NLSY project. One of the main reasons why we selected the Accurint database for our evaluation was the success NORC had in managing to keep attrition low by using Accurint as the locator source for the people in the NLSY sample. Our study used computer-assisted clerical matching to find records for the NLSY movers in the Accurint database that corresponded to the addresses where they lived since their previous interview.

Initially, we sampled 2,999 NLSY respondents, which resulted in 4,105 moves over three rounds of the NLSY (combining moves to the interview address and intervening moves since the last interview). NLSY had the street address for the moves to the current address since that is where almost all the in-person interviews occurred. However, NLSY collected only city, state, and zip code for the intervening moves since the previous interview.

We found the results of the matching between the NLSY respondents and Accurint varied by the amount of agreement between the records that we required. When matching on name only, 90 percent of the respondents were found in Accurint. After adding the requirement of matching on state, the match rate dropped to 69 percent. When matching on name, city, state, and 5-digit Zip code, the match rate dropped again to 56 percent. The decline in the match rate suggested that the erroneous match rate might be unacceptably high in the less restricted matches.

For the matched moves with agreement on name, city, state, and 5-digit Zip code, the difference in the move dates for Accurint and NLSY was much larger than we expected. This result made us skeptical that the linked records represented the same move event. Therefore, we decided to match movers at the street address where the interview occurred. We had nearly 6,000 interview addresses for our NLSY respondents. Our match rate in the second match attempt was very similar to the previous round of matching – it was 57 percent, but we were confident that we had the correct address this time. However, only 19 percent of the original interview addresses had move dates available in both the NLSY interview and Accurint. We think the reasons may be that the Accurint database does not cover certain populations well, or that some moves did not generate a record that Accurint would acquire, such as utility bills, credit cards, etc.

Now, we turn out attention to our NCOA/Census Study where we started in a different way than the Accurint/NLSY study. In this study, we started with the administrative record NCOA file and then conducted the interviews. We selected 13,500 housing units which submitted a change of address for March or April 2010 and for which we could find telephone numbers in a third-party database. The U.S. Census Bureau conducted a telephone interview that collected the address of the residence, all the people living in the residence, and other places where they lived during the year, including move dates. The study used clerical matching between the NCOA records and the survey interviews.

The NCOA/Census Study achieved a 66 percent response rate overall using AAPOR Response Rate 2 (American Association of Public Opinion Research 2011) that includes sample units of unknown eligibility in the denominator. However, when we looked at the data more closely, we found only about 25 percent of the original sample reported a name and that matched the name and forwarding address on the NCOA file. Furthermore, only 15 percent of the original sample appeared to be the same household and reported a move.

We think there are two possible contributors to the low percentage that appear to be the same household and report a move. One is that possibly the telephone number we had for the household was not correct. Either the linking to find a telephone number for the forwarding address found a number for a previous resident or the link was correct but the person with the name on the NCOA form moved to another address before our study called. The other problem is that the survey respondent did not report a move, which occurred either because the person forgot to report the move or because the NCOA filing did not correspond to a move.

4. Locating Same Event

The events that we were studying were residential moves. Although moving is not typically a frequently occurring event, the U.S. Census Bureau estimates that about 12% of the U.S. population moves every year U.S. Census Bureau, 2011. In both our studies, we found wide variability between the two sources as far as the move date.

For the NLSY/Accurint study, when we examined the difference between the Accurint start date for the address and the NLSY report of the date of the move to the address, we saw large discrepancies. Sometimes, the Accurint start date was before the interview, and sometimes the start date was after the interview. For example, some people in the NLSY said they moved to the address in the same month as the interview, but linked to an Accurint record 50 months earlier. We concluded we had the wrong event. Otherwise, it would suggest that Accurint knows where a person is going to move four years in advance. One extreme example occurred for two respondents who said they moved to the address four months earlier, but linked to Accurint records about 25 years earlier. Remember the NLSY people were between 23-29 years old so the more likely hypothesis is that their Accurint records were associated with these people's birth. To our knowledge, Accurint does not perform any editing or correction of dates that would cause the discrepancies we observed.

Some of the links with the wide differences were parental addresses, but most were not. To attempt to get the correct event, we had to further subset our data and only looked at matches within a particular time-frame of each other (e.g., 1 to 2 years), then we examined how results changed as that time frame changed.

In the NCOA/Census study, getting the correct event was more straightforward. In the survey, we collected dates and had almost 2,000 move dates reported. Around 200 of them were for situations where the person reported moving to the forwarding address, but the person did not report the month of the move to the address. Without the date of the move, those situations could not be considered in our analysis.

5. Generalizing Results

As you can imagine, our sub-setting of data affected how much we could generalize our results. However, our analyses in both studies, limited though they were, did find greater increase in memory error of move dates starting somewhere between 6 and 10 months following the move.

For the NLSY/Accurint study, our respondents were 23-29 years of age because of who was in the NLSY cohort. Even though we started with nearly 3,000 respondents, in the end we could use only 410 move events for analysis. These were events where we could be certain we had the correct person, address, and move. We were unable to find adequate controls for weighting movers in this age group, which made the issue of generalization mute.

The NCOA/Census study started with a dataset that contained movers, those who file a change of address. However, NCOA does not include all movers (because not all movers file a change of address with the USPS) and not everyone who files a change of address is a mover (one can conceive of reasons why you might forward your mail when not physically moving). We started with 13,500 sampled households and ended with 1,740 move events. Again, we did not have adequate controls to weight movers in March/April of 2010. The historical data we could find did not seem appropriate for weighting our data.

6. Frequency of Record Updates

The timing of the updating of the administrative or third-party data had an impact on our analysis.

The flow for the NLSY/Accurint study was that the NLSY interviews we used occurred from October 2006 through May of 2009. NORC pulled the Accurint data in 2011 for these cases and then the matching was done. We did not think too much about the timing of our evaluation, but in hindsight, this worked really well. Doing the search in Accurint so long after the interviews gave it enough time to be updated and enabled finding a match for some of the events that entered Accurint months and even years after the interview. If we had done the evaluation closer in time

to the interview, that third-party database would not have had as many updates and records added with the implication that even fewer cases would have been available for inclusion in our analysis.

In the NCOA/Census study, the timing of the administrative records use in the NCOA/Census study perhaps had an effect on the response rates. The NCOA records for March and April 2010 were pulled and then the forwarding addresses were matched to telephone numbers from a different third-party administrative database. Using those telephone numbers, we then conducted the interviews in June, September and February. We saw a 69 percent response rate in June, but it dropped to 63 percent by February. We think the drop was because the addresses were matched for telephone numbers in May, and some of the numbers were out of date by the following February.

When doing these types of studies, one has to remember that databases error. Potential sources of database error include the following:

- An individual or household reporting their own information may be higher quality than information coming from a magazine or utility (the source of some Accurint records);
- Differences in definition, such as a temporary address versus a permanent address;
- A legal residence versus usual residence (where the person usually lives); and
- The quality of some information may be more important to the database owner than other types. For example, birth date is very important to the U.S. Social Security Administration since a person's age determines eligibility for the benefits it provides.

Therefore, a researcher has to be cognizant of the implications of the data sources, updating patterns, and data priorities when selecting a database for research on survey reporting error in the variable of interest. A database may be a good source for some types of variables in some months, but not all year.

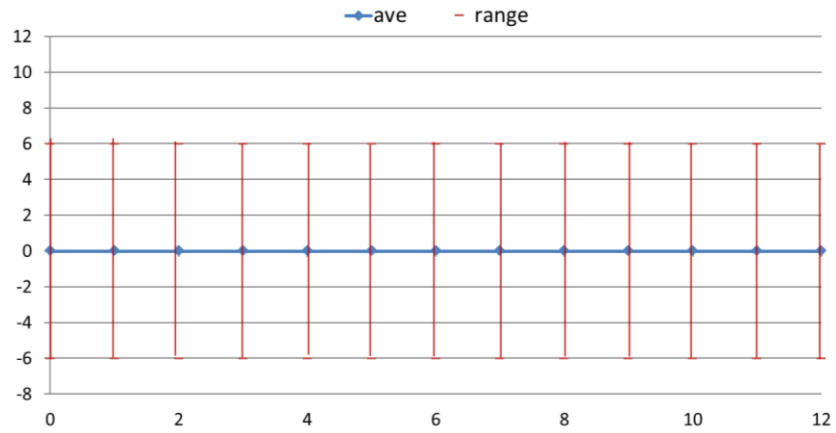
In studies where the database provides the frame for sampling movers, the updating of the database affects the study at the sample selection phase in that there may be undercoverage of movers as in the Census/NCOA study. In the case of the NCOA file, overcoverage may be present since some changes of address may not correspond to moves.

When the database is not the sampling frame, but rather the source of comparison for the survey report of the move as was the case in the NLSY/Accurint study, the issues surrounding updating of the database impact the presence of the event in the database. Below we illustrate two patterns for errors in databases using errors in addresses for movers – namely on when a database acquire the new address associated with a move. Addresses for movers are probably one of the most error-prone types of data.

Figure 6.1 contains an illustration of a database that is continuously acquiring data and updates monthly. This illustration assumes that a new address is used at most 6 months prior to the physical move. For this database, the average of the length of time to acquire a new address is zero – or the month of the move. However, there is a wide range in the number of months that it takes the database to acquire the address.

We think Accurint may have a pattern similar to the one shown in Figure 6.1. We did not detect a bias in length of time for Accurint to acquire a new address, but we had a small amount of data and made assumptions in our analyses. Having an average acquisition time that is stable over time may make a database a reasonable candidate for evaluating reporting error when the data is collected over a period of time, such as the three-year period in our NLSY/Accurint study. An additional consideration is whether stability in a database's overall acquisition time also exists for the subpopulation of interest and the data used in the analysis.

Figure 6.1.
Illustration: Length of time for mover's new address to appear in database that acquires records from many sources each month

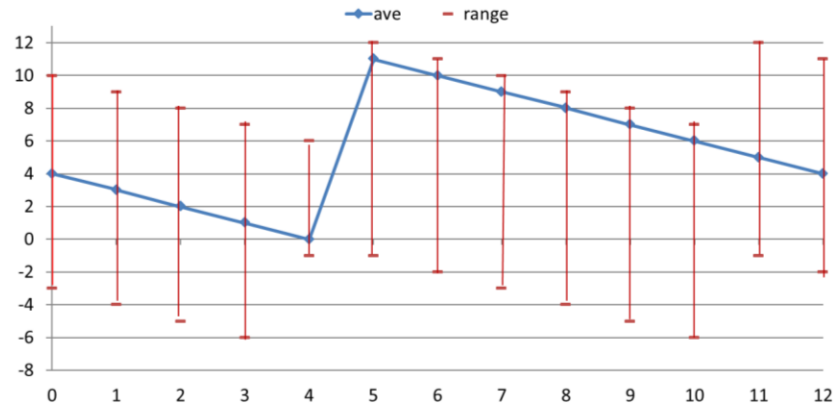


Assumptions: Database is continuously acquiring new records.
A new address is used at most 6 months prior to move and at latest, is updated 6 months after the move

In contrast, Figure 6.2 contains the second illustration of another type of database, one that updates once a year in Month 4 but also collects a few updates in Month 10. The main idea is that the database is updated yearly, not monthly. Again, this figure reflects a conjecture about the pattern of acquiring a new address for such a database. Figure 6.2 shows that Month 4, the update month, has the lowest average length of time for an address to get into the database. The range of the time to acquire the address is not symmetrical about the average while the error in the database in Figure 6.1 was symmetrical.

The database Figure 6.2 with once-a-year updates has a very different error pattern for movers' addresses than the previous database with a continuous acquisition of information and monthly updates. This database's important feature is that the length of time to acquire a new address varies during the year so it is not a good candidate for evaluating survey reports of move month. However, it possibly could be a good candidate for evaluating other variables, such as a person's address in Month 4 or variables of an annual nature, such as annual income.

Figure 6.2.
Illustration: Length of time for mover's new address to appear in database that acquires almost all its records in Month 4



Assumptions: A few records acquired in Month 10 but very few other times.
A new address is used at most 6 months prior to move.

7. Summary

When planning a study, the bias and random error in the database for your variable of interest is an important consideration. The database error properties may affect:

- The timing of the survey interviews, such as conducting interviews as close to database updates as possible.
- The methodology you choose for your analysis.

In summary, we learned several things from our two experiences that are applicable in other studies that use administrative records or third-party databases to evaluate error in survey reports:

- Plan the matching carefully, especially for frequently occurring events
- Know the definitions and limitations of all data
- Plan the timing of the data pulls optimally
- Be prepared for a much smaller dataset than expected
- Be aware of how bias and random error in your chosen database affect your population of interest, your study design and the ability to generalize your results.

References

American Association for Public Opinion Research (2011), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition*. AAPOR. http://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/StandardDefinitions2011_1.pdf (Accessed February 6, 2015).

Belli, R. F. (1998), The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory, 6*, 383-406.

Johnson, D. S., C. Massey, and A. O'Hara (2015), "The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility," *The ANNALS of the American Academy of Political and Social Science January 2015*, 657. pp. 247-264.

Krishnamurty, P. (2012), "Memory Recall of Migration Dates in the National Longitudinal Survey of Youth, 1997 Cohort: Statistical Analysis and Evaluation" (Census R&D 2014 IDIQ Contract No. YA 1323-09-CQ-0053). NORC.

Linse, K., T. Pape, L. Rosenberger, and G. Contreras (2010), Census Coverage Measurement Survey Recall Bias Study. U.S. Department of Commerce, U.S. Census Bureau, Washington, DC.

Mulry, M. H. (2014), Measuring Undercounts for Hard-to-Survey Groups. In R. Tourangeau, N. Bates, B. Edwards, T. Johnson, and K. Wolter (Eds.), *Hard-to-Survey Populations.* (Chapter 3). Cambridge University Press, Cambridge, England. pp. 37 – 57.

Mulry, M. H., E. M. Nichols, and J. H. Childs (2014), "Study of error in survey reports of move month using the U.S. Postal Service Change of Address records. In *JSM Proceedings*, Survey Research Methods Section. American Statistical Association, Alexandria.

Mulry, M. H., E. M. Nichols, and J. H. Childs (2015), "Evaluating recall error in survey reports of move dates through a comparison with records in a commercial database." Unpublished manuscript. U.S. Census Bureau. Washington, DC.

Nichols, E. M., M. H. Mulry, and J. H. Childs (*Forthcoming in 2017*), "Using administrative records data at the U.S. Census Bureau: Lessons learned from two research projects evaluating survey data." In Biemer, P.P, Eckman, S., Edwards, B., Lyberg, L., Tucker, C., de Leeuw, E., Kreuter, F., and West, B.T. (Eds.), *Total Survey Error in Practice*. Wiley. New York.

O'Hara, A. (2014), "Comments on: Laying the foundations for a new approach to Census taking in Ireland by John Dunne and Steve MacFeely." *Statistical Journal of the IAOS, 30.* pp. 367-368.

U.S. Census Bureau (2011), *Mover Rate Reaches Record Low, Census Bureau Reports*. Last accessed August 28, 2012: http://www.census.gov/newsroom/releases/archives/mobility_of_the_population/cb11-193.html

U.S. Postal Service, NCOA^{Link} (2015), [WWW document]. Last accessed on August 7, 2015: https://ribbs.usps.gov/index.cfm?page=ncoalink