Discussion of Utilizing Administrative Records and Adaptive Design in the 2020 Census

Michael D. Larsen[1]
[1]Department of Statistics, George Washington University, Washington DC 20052

Abstract
This is a discussion of the three papers presented in the invited session "Utilizing Administrative Records and Adaptive Design in the 2020 Census" in the Government Statistics Section at the annual meeting of the American Statistical Association. The three papers together focus on efforts to improve methods for non response follow-up after the initial stages of self-response in the 2020 Census. The discussion presents some background on the rationale for the research and argues that a sustained effort is needed to aim for the greatest benefit from use of administrative records. Comments are given on the three papers. In general, the three papers have demonstrative practical approaches for administrative record files to be used to decrease cost and maintain quality during the non response follow-up of the U.S. Decennial Census.

Key Words: Decennial Census, Non response follow up

1. Use of Administrative Records in Decennial NRFU

This is a discussion of the three papers presented in the invited session "Utilizing Administrative Records and Adaptive Design in the 2020 Census" in the Government Statistics Section at the annual meeting of the American Statistical Association. *The views expressed here are those of the author and do not necessarily reflect those of the others involved in the session or any organization.* The three papers in the session were

- Administrative Records Research to Reduce Contacts in the 2020 Census (Morris, Keller, and Clark 2015), presented by Morris,
- Adaptive Design Research for the 2020 Census (Konicki and Adams 2015), presented by Konicki (2015), and
- Imputation Research for the 2020 Census (Keller 2015), presented by Keller.

The three papers in this session together focus on efforts to improve methods for non response follow-up (NRFU) after the initial stages of self-response in the 2020 Census (the 2020 Decennial Census of the United States of America).

The 2010 U.S. Census was the most expensive ever in total cost and cost per household. The 2020 U.S. Census must hold costs down. Administrative record (AR) files, such as Internal Revenue Service (IRS) tax records, Center for Medicare and Medicaid Services

(CMS) health records, U.S. Postal Service (USPS) delivery files, and Social Security Administration (SSA) records, cover most of the U.S. population. That is, most members of the population are included in one or more of these files. The AR files, however, have limitations that preclude them from being a direct replacement of the decennial census enumeration. These AR files

- Do not cover everyone,
- Are not necessarily provided with an address that is geographically linkable to Census address file,
- Are not necessarily dated to or current on Census day (April 1, 2020), and
- Contain some errors and are not necessarily unique.

Despite these limitations, the AR files are worth exploring for targeted uses to improve 2020 Census enumeration while reducing costs.

The overall research question for AR files in the 2020 Census is

- Can a system of administrative records be assembled and maintained for the purposes of **reducing cost** and **increasing quality** in one or more aspects of the decennial census (e.g., NRFU)?
- How can one **validate** decisions and procedures involved in the use of AR files in the decennial census?

A potential plan to use AR files in the decennial census can increase its chance for acceptance and success if it

- Uses select records of high quality,
- Uses records in appropriate areas where the records are highly accurate,
- Uses the files for appropriate purposes as given in a clear operational rule, and
- Does not try to be a panacea for all difficulties in the decennial census.

The papers in this session have shown that there is potentially great benefit to selected uses of some AR files. Further work is needed to refine the operational rules for using AR files in NRFU, formulate a rationale for adopting the operational rules, and working out details of the plans.

The first question of the overall research statement mentions **assembling** and **maintaining** a system of AR files for use in the decennial census. Early every decade a new project is proposed to study the availability of administrative lists, research their content and usability for population enumeration, devise methods for utilizing the AR files in some aspect of the decennial census, and consider legal and logistic issues. For example, the author was exposed to issues of census enumeration before Census 2000 (Larsen 1999abc). Instead of starting and then stopping AR research efforts with the build-up to and then reduction after the decennial census, a continuous effort over time is needed. Ideally the question of how to gather records from various sources and how to process files of different types would be solved and the knowledge of how do to so would be transmitted within the U.S. Census Bureau across time. Legal agreements that allow use of AR files in yearly operations should be framed, debated, and approved. Validation tests of use of AR files for various purposes should be conducted every year in conjunction with ongoing major surveys, such as the American Community Survey

(ACS). The knowledge and experience gained through a sustained effort will enable the Census Bureau to aim for the greatest benefit from use of administrative records in the decennial census and in its ongoing large-scale survey operations.

As an aside, the National Academies of Sciences (NAS) Panel on Census 2010 (2009-2014) examined issues related to costs and accuracy in the 2010 Census and made several recommendations looking forward to 2020 (National Research Council 2011). Topics included the use of administrative records, self response modes including the Internet, enhancements to geographical information systems, and whole-scale enterprise/process engineering. It was an interdisciplinary team that was able to interact extensively with experts in many divisions of the U.S. Census Bureau. A new NAS Standing Committee on Reengineering Census Operations (2015-2018) is continuing detailed discusses with representatives of the U.S. Census Bureau. Session organizer Tom Mule, the speakers in this session, and the discussant/author of this paper participate in these discussions. See also National Research Council (2010, 2004a, 2004b, 1994) and other reports.

Sections 2, 3, and 4 make specific comments on the three papers presented in this session. Section 2 concerns "Administrative Record Research to Reduce Contacts in the 2020 Census" presented by Darcy Steeg Morris. Section 3 discuses "Adaptive Design Research for the 2020 Census" by Scott Konicki. Section 4 focuses on "Imputation Research for the 2020 Census" from Andrew Keller. As has been mentioned, the focus of the session and these three papers is research on the use of a select collection of administrative records in Non Response Follow-up Operations in the 2020 U.S. Decennial Census.

Some details in the following discussions are explained in greater detail in the source papers from the conference presenters. Please see those papers for definitions and further information.

2. Morris: Reduce Contacts

The paper by Morris, Keller, and Clark (MKC 2015) considers

- the replacement of part of the vacant/delete housing check (one step of NRFU) with evidence from administrative records, and
- the replacement of part of the in-person NRFU interview with evidence from administrative records.

In the second case, if a unit is judged to be occupied, AR files could be used to inform the estimated household count and its composition.

The authors of this paper, in conjunction with the authors of the other two papers, use processes that create counts by household based on multiple AR files. If there are multiple records from AR sources at a given address, then choices must be made concerning which records are judged to be accurate. This is an important step in this research. The papers do not concern the details of this process, and comments are not presented here. Needless to say, handling multiple records per housing address is an important topic.

The authors use the 2010 Census as a test-bed for evaluating proposals for using AR files. Information, including administrative records, self-response records, and NRFU

interview records, that would have been current at the time of the 2010 Census are available to the authors. MKC (2015) fit multinomial logistic regression models with constrained optimization incorporating costs of decisions and errors in classification to predict whether a household that does not answer the initial part of the census should be classified as delete, vacant, or occupied. They have person-level and household-level models which they combine for a composite answer.

Results so far are at national level. The authors have addressed many complications, such as proxy response in 2010 and the difference between households versus persons in NRFU. Results for one scenario are explored in detail. It appears possible to use administrative records to remove more than 10% of NRFU case load without adversely affecting the population count. A dollar savings is not attached to this, but presumably the savings would be substantial (e.g., $200 million?).

A few questions can be asked about the plan for ongoing research and development.

- A decision rule procedure is needed for 2020. A question that must be answered is, based on this research, what can be advocated as an operational plan for 2020?
- How do error rates for NRFU scenarios utilizing AR files compare to 2010 Census errors that were observed for proxy responses, which were often collected after several failed attempts to reach the target interviewee, and early versus late interviews in NRFU?
- Results are sensitive to constraints and costs. It would be interesting to learn if this sensitivity is different by region or factor such as degree of urbanicity.
- Creating variables for use in this study required some record linkage. Has the record linkage been validated and otherwise evaluated? That is, how clear is it that individuals on IRS records at other locations are different people?
- Can additional interactions be incorporated into regression models? One interaction could be a code if an individual appears at an address on two or more administrative records. Coding this interaction into a single variable, in addition to indicator variables for presence on the separate administrative files, could strengthen model predictions for a subset of persons.
- The household category models rely partially on the existence of children in administrative records. How often do children in a household show up in administrative records used in this study? If children do not show up very frequently, then prediction could be compromised.

How can one improve prediction for children and for the poor when using administrative records? Some programs, such as

- The USDA's Supplemental Nutrition Assistance Programs (SNAP),
- HHS's Temporary Assistance for Needy Families (TANF) program,
- The USDA's Women, Infants, and Children (WIC) program,
- The USDA's National School Lunch Program (NSLP) and School Breakfast Program (SBP), and
- Unemployment insurance,

should provide partial coverage of some populations at risk of not being enumerated and of not being on other AR files. There is not much experience, to the discussant's knowledge, of acquiring and working with data from these programs. Some of these

programs are state-based programs, and data reporting to the federal government is not completely standardized. As a result, bringing additional files into the collection of administrative records used to reduce cost and effort for NRFU will require additional research.

In conclusion, the research presented in this paper suggests that at a minimum NRFU workload can be reduced by 10% through the use of administrative records without compromising quality at a national level. Some choices made in development and applying models were conservative and it is possible that gains could be higher if these efforts are pursued further.

3. Konicki: Adaptive Design

In the 2010 Census, NRFU involved 3 in-person visits and 3 telephone calls for households with no mail response. This was followed by an attempted proxy interview. An adaptive contact strategy is anticipated in the 2020 Census. In some places, fewer visits and calls will be made before resorting to imputation or other methods. Stopping earlier in select areas should reduce costs while maintaining quality.

The measure of quality used in this paper by Konicki and Adams (2015) is the rate of proxy interviews. A small proxy rate is good. Assuming an average of three visits per household, modifications of the interviewing schedule are chosen to minimize the variance of proxy rates across block groups in NRFU. A small variance means that proxy rates are quite similar across blocks. Areas with good response to the initial phase of census data collection and to early NRFU efforts then will be able to reduce efforts. Areas with poor response will need additional effort to reduce proxy rates. Is variance of proxy rates the only standard for success? Presumably higher variance would be accepted if the overall rate were lower, or if rates were low and uniform in 90% of the areas, but high and outlying in others. One might then address those difficult areas in another manner.

The paper also suggests controlling the day and time of visits to increase response rates. Models are fit to American Community Survey (ACS) contact data (2012-13). The modeling of contact success by day and time makes sense but predictors are weak. Possibly one needs to incorporate a failed attempt into prediction for second attempt. In general one would want to incorporate information on all failed attempts to predict the best day and time for the next attempt. At some point evidence might accumulate to suggest a very low probability of a successful interview on any day or at any time. Assuming a heuristic rule is created for distributing contact attempts. One then would want to evaluate the performance of the rules.

The paper describes Census tests from 2013, 2014, and, now, 2015. An initiative called *ROCkIT* at the Census Bureau seeks to reengineer how enumerators are assigned work, how enumerators communicate and are managed, how optimal routes for travel are computed, implemented, and monitored, and priority for visiting cases. As such, it seems that *ROCkIT* could potentially benefit from coordination with this research effort. Similarly, this research needs to be integrated with general field reengineering.

The paper uses 2015 Census test results from Maricopa, AZ, to study computational methods for optimal allocation determinations. The results seem plausible. It seems reasonable to think about practical tests in the field.

Goals of this work are laudable. Efforts like this could lead to significant reductions in cost *and* improvements in quality. The number of contacts and contact times need to be coordinated with *ROCkIT* tests.

One experiment that one can imagine would be to collect data from up to six household visits plus a proxy respondent in some areas. Then you could compare the counts for a number of plans. Plans that could be compared would be (1) use administrative records to fill in households, (2) allow *k* (*k=1* to *6*) visit(s) to a NRFU household and then utilize AR, or (3) allow *k* visit(s) and a proxy response. This idea involves duplicate data collection, but it would enable direct comparison of content of interviews, administrative records, and proxy responses. Sample size, funding, and the time frame before Census 2020 probably make a test such as this unlikely to occur, but it is worth bringing into the discussion.

4. Keller: Imputation Research

The paper by Keller (2015) discusses the overall Census process, including the sequence of mail contact and reporting, NRFU, and then imputation of missing or unknown characteristics. The 2020 Census envisions less NRFU, so (much) more imputation could (will) be needed. Imputation and proxy response has been used for decades; this research seeks to utilize better data sources and make choices that reduce cost while maintaining quality. The paper uses administrative records to build models that predict housing unit *status*, the housing unit person *count* in occupied units, and also person *characteristics*.

The procedure begins with vacant/delete and occupied housing models. These models are similar to those in the paper by Morris, Keller, and Clark (2015). Counts then are imputed based on composite administrative records for occupied housing units. People with unknown characteristics are replaced with some AR records and/or some imputation of characteristics is performed. Some distinctions in methods sound subtle, but they have real implications for operations and data quality. The method is tested using information from the 2010 Census and its NRFU operations. This is an excellent test bed for these methods.

A number of comments have been made previously that are relevant to this research. Those comments will not be repeated here.

It appears that the procedure suggested here has great potential, based on the study using 2010 Census data, for reducing AR visits while maintaining data quality. Results now are at the national level. It will be interesting to look at distributional results at smaller levels of aggregation and for demographic groups. One could consider fitting sub-national models for prediction. These models might vary by region, but they might vary more by other factors. If counties are grouped by factors such as urbanicity and employment/poverty levels, then one might find different models for the vacant/delete/occupied classification are useful. One might find that some AR data sources and some modeling strategies work well in some parts of the country, but not others. Finding a plan that is workable operationally, defendable in terms of cost savings and impact on quality, and not overly sensitive to assumptions will be challenging. The research of this and the other two papers suggests that it could be a reachable goal.

5. Summary

Three interesting and practical papers demonstrating potential for use of AR files in 2020 Census NRFU operations have been presented. The next steps are critical for successful utilization of administrative records in the 2020 Census. It is necessary now to decide operational rules, test aspects of those rules to the degree possible, and evaluate their use in the field. This needs to be done within the framework of a full census system, or as close to one as is feasible before the actual 2020 operations are underway. Results of ongoing developments and testing should be interesting to see.

While preparing for 2020, it simultaneously is important to look ahead to 2030. Formulating a plan to test in 2020 the use of programs data on the poor and involving children for the purpose of improving enumeration of said groups could have important implications for developments following the 2020 Census. A sustained effort to assemble and maintain AR files, or at least the experience of how to obtain, process, and use various files, should not be postponed. The potential to reduce cost and improve quality in the decennial census and in continuous large-scale surveys through the use of AR files is too important to not make an adequate effort.

Acknowledgements

The author wishes to thank the organizer, Vincent (Tom) Thomas Mule of the U.S. Census Bureau for organizing the session and inviting me to discuss three interesting papers. Thanks also to the speakers for preparing their papers and presentations well in advance of the session.

References

Keller, A. (2015). Imputation Research for the 2020 Census. Proceedings of the Government Statistics Section, American Statistical Association: Alexandria, VA.

Konicki, S., and Adams, T. (2015). Adaptive Design Research for the 2020 Census. Proceedings of the Government Statistics Section, American Statistical Association: Alexandria, VA.

Larsen, M.D. (1999a). Identifying residents in administrative records that do not match the Census. *Turning Administrative Systems Into Information Systems.* Internal Revenue Service, Statistics of Income Publication 1299.

Larsen, M.D. (1999b). Identifying residents in administrative records that do not match the Census. Proceedings of the Government Statistics Section, American Statistical Association: Alexandria, VA.

Larsen, M.D. (1999c). Predicting the residency status for administrative records that do not match census records. *Administrative Records Research Memorandum Series*, #20, Census Bureau, U.S. Department of Commerce.

Morris, D.S., Keller, A., and Clark, B. (2015). An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census. Proceedings of the Government Statistics Section, American Statistical Association: Alexandria, VA.

National Research Council. (1994). *Counting People in the Information Age*. Washington, DC: The National Academies Press.

National Research Council. (2004). *Reengineering the 2010 Census: Risks and Challenges*. Washington, DC: The National Academies Press.

National Research Council. (2004). *The 2000 Census: Counting Under Adversity*. Washington, DC: The National Academies Press.
National Research Council. (2010). *Envisioning the 2020 Census*. Washington, DC: The National Academies Press.
National Research Council. (2011). *Change and the 2020 Census: Not Whether But How*. Washington, DC: The National Academies Press.