Big Data and Censuses:
The Challenge for the 2020 World Round
# of Population and Housing Censuses

Talking points developed for the
Population and Housing Census Lunchtime Seminar

46th Session of the
United Nations Statistical Commission
United Nations Headquarters, New York
3 March 2015

Lisa M. Blumerman
Assistant Director, Decennial Census Programs
U.S. Census Bureau

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Good afternoon. I'd like to begin by thanking the Chair and Organizer of this session for the opportunity to present this afternoon, to share the United States experiences with big data as we prepare for our 2020 Census, and to learn from others as to their experiences in this area. The panel was asked to address three issues today:

- What is the main focus in terms of using big data for a population and housing census?
- Is the primary focus on governmental big data, corporate big data, or both?
- What are the main findings so far?

These are exciting questions for me to consider and discuss as the plans for the United States 2020 Census unfolds.

To begin, please allow me to provide a little context on the design for our 2020 Census.

A principal goal for the U.S. 2020 Census is to conduct the Census at a lower cost than our previous 2010 Census while maintaining high quality results. In order to do this, we are reengineering the majority of our Census processes, including our data collection techniques, our methodologies, and our field structure. As we have planned over the past 5 years, we have focused our research and testing efforts in the areas with the greatest opportunity for cost savings and the introduction of innovations, including the use of big data and tools to manage big data.

As part of our redesign, the United States has several high-level guiding principles that we are operating under. The first guiding principal for the design of the 2020 Census is to:

1) Use the Internet to increase self response; [optimizing self response]
    - Internet data collection
    - Notify Me – early engagement for people to provide an email address or mobile phone number that we would use to contact them when we are ready to begin data collection.
    - Advertising, partnership, and promotion

- Non-ID processing enables processing of a questionnaire without an identification number, for those households that did not receive a notice or questionnaire mailed to their housing unit.

2) Automate operations to increase productivity and reduce temporary staff and offices; [reengineering field operations]
3) Use information people have already given the government for other purposes that can be used to answer Census questions and reduce the follow-up workload; [reengineering nonresponse follow-up] and
4) Update existing maps and addresses to reflect changes rather than walking every block in every neighborhood in America in 2019. [reengineered address canvassing]

Our work today in these areas allows us to estimate that we will achieve a savings of approximately $5.1 billion, as compared with our 2010 Census in 2020 dollars.

The use of big data isn't new for the United States. We've been using administrative data, such as tax data, for decades to improve our data collections. Today, there is a new generation of big data – as the electronic environment flourishes – that we must keep up with. We are researching ways to utilize these new data sources in our collections in order to increase efficiencies and to reduce costs and the time it takes to disseminate statistics. At the same time, we must also continue to maintain the quality of the official statistics.

With that as a brief background, I would like to delve into the three questions posed and discuss the United States' proposed use of big data in the context of two of our innovation areas and operations – the address canvassing operation and the nonresponse follow-up operation.

A specific new application of the use of big data proposed for the 2020 Census is in our plans to reengineer our address canvassing operation. Traditionally, in the years prior to the decennial census, the Census Bureau employed a large-scale field collection effort to walk every block[1] in the United States (about 6.7 million blocks and drove 137 million miles). The availability of high quality, high resolution satellite, aerial, and street-level imagery, with increasing frequency of updating, now provides a viable and effective alternative to fieldwork for many parts of the United States, Puerto Rico, and the Island Areas[2]. For the 2020 Census, we have redefined address canvassing to be a combination of in-office and in-field canvassing where we will continue to canvass every block, but we will only conduct in-field, on the ground canvassing where it is necessary.

The goal of reengineering address canvassing is to eliminate a nationwide in-field address canvassing in 2019. Instead, the Census Bureau proposes to utilize the use of statistical models to help predict where change is occurring, combined with aerial imagery and change detection techniques to identify the areas where in-field canvassing is required.

We are working now with terabytes worth of data from federal, state and local government sources, as well as the private sector to update our Master Address File (MAF) and determine where in-field work is required. Specific federal files we are working with include: extracts from the United States Postal Service, Indian Health Service Registration File, U.S. Department of Housing and Urban Development Public and Indian Housing Information File, and the Selective Service Registration File. In addition to the files from the Federal Government, we also maintain an extensive partnership program with state and local governments where we have been receiving “partner” files for several years and have processed and ingested those files into

[1] Blocks (Census Blocks) are statistical areas bounded by visible features, such as streets, roads, streams, and railroad tracks, and by nonvisible boundaries, such as selected property lines and city, township, school district, and county limits and short line-of-sight extensions of streets and roads. Generally, census blocks are small in area; for example, a block in a city bounded on all sides by streets. Census blocks in suburban and rural areas may be large, and irregular, and bounded by a variety of features, such as roads, streams, and transmission lines. In remote areas, census blocks may encompass hundreds of square miles. Census blocks cover the entire territory of the United States. Census blocks nest within all other tabulated census geographic entities and are the basis for all tabulated data.

[2] The Island Areas of the United States are American Samoa, Guam, the Commonwealth of the Northern Mariana Islands (Northern Mariana Islands), and the United States Virgin Islands. Sometimes the Island Areas are referred to as "Island Territories" or "Insular Areas." For the 1990 and previous censuses, the U.S. Census Bureau referred to the entities as "Outlying Areas."

our address frame. We are also using private sector data where we have found gaps that exist in the governmental data.

For the purpose of address canvassing, a large focus of our efforts with big data include both governmental and commercial (corporate data). We are looking at the best methodologies to integrate these data, predict change, and then take appropriate actions by either updating our Master Address File or conducting an in-field canvass.

Current research in this area is very promising. We just concluded our first test of the new approach, called the Address Validation Test, and we are in the data analysis stage now. In this test, we developed two independent statistical models that predict the stability of a block, and then conducted a dependent listing operation of 10,100 blocks in the United States. The result of that dependent listing is now being compared to the output from the statistical models to assess the veracity of the models. A second component of this test identified 29 counties in the United States for which we had aerial imagery in-house that overlapped with the 10,100 blocks. Both automated and manual change detection techniques were utilized in those 29 counties to identify blocks with change and of those approximately 700 blocks were identified as suitable for an in-field canvass. The results of this in-field canvass will be compared with the results of the change detection work, as well as the results of the dependent listing operation. We expect to have these findings by early April and they will inform the proposed methodology for the 2020 Census.

A second application of the use of big data in the 2020 Census design is in our planning and operationalization of our nonresponse follow-up operation. The reengineering of the components of this operation encompasses two of the Census Bureau's innovation areas: Utilizing Administrative Data and Reengineering Field Operations.

The goal of Utilizing Administrative Data is to use data that the public has already provided to the government, and, potentially, third-party (or commercially available) data to reduce the nonresponse follow-up (NRFU) workload. The Census Bureau proposes to use data from internal and external sources, such as the 2010 Census, the United States Postal Service (USPS), and the Internal Revenue Service (IRS) to identify vacant housing units and those units that do not meet the Census Bureau's definition of a housing unit (deletes). The data sources may also be used to enumerate the population in cases of nonresponse.

During the 2010 Census, the nonresponse follow-up universe included 50 million housing units. Each of those units received at least one personal visit, resulting in the identification of 31 million occupied and 14 million vacant housing units. Another five million units were deleted because they did not meet the Census Bureau's definition of a housing unit[3]. Vacant and deleted units accounted for about 38 percent of the non-responding universe. The use of administrative data to avoid the expense of conducting a personal visit to an address to discover vacant units for which no questionnaire was, or could be returned from, is one of the key cost drivers of the Census. By using administrative data to identify these vacant units, we can eliminate and reduce the field effort involved.

In terms of our initial research and testing in this area, our primary focus has been on the use of government provided files (both federal and state level files). We have supplemented our research and testing with commercial files, as necessary, when they could supplement the existing federal data. We have conducted two field tests of these new methodologies and will be beginning our third field test in March of this year. To date, the research has been promising and has informed changes in the methodologies employed in each round of testing. For example, as a result of the 2013 Census Test, we identified additional information in the administrative data available from the Postal Service that informed the identification of vacant housing units in the 2014 Census Test.

[3] The U.S. Census Bureau definition of a housing unit is a house, an apartment, a mobile home, a group of rooms, or a single room that is occupied (or if vacant, is intended for occupancy) as separate living quarters. Separate living quarters are those in which the occupants live and eat separately from any other persons in the building and which have direct access from the outside of the building or through a common hall

A second component in our efforts to reduce the cost of our nonresponse follow-up operation is our reengineering of our field operation. The goal of Reengineering Field Operations is to use technology to more efficiently and effectively manage the 2020 Census fieldwork. The Census Bureau plans to develop an operational control system that manages tasks and makes decisions typically made by humans (e.g., case assignments, contact attempts). Additional modernization includes a streamlined approach to implementing and managing field operations through a new field management structure, including the infrastructure, field staff roles, work schedule, and staffing ratios.

Specifically, we are utilizing big data to help us in the efficient assignment of work. Our operational control system (case management system) will consider outstanding work, enumerator home location, and enumerator attributes to tailor a work assignment that is appropriate for each enumerator on a daily basis. Enumerator assistance from remote operations centers also reduces the enumerator-supervisor ratio.

We will be using real-time paradata to manage the work and our enumerators by providing alerts. Using real-time data within our control system will alert supervisory level employees when specified performance or conduct issues are identified. This could be as simple as the employee completed cases for the day but did not send us their corresponding payroll and expense information, or as complex as an employee's device location being nowhere near where we believe the work to be.

Behind the scenes, a significant component of this effort, and the use of big data, is in the development of the Constructive Simulation Model. This is the model that is at the core of our control system and is a three-dimensional model that utilizes 2010 Census data, data from the American Community Survey[4], and real-time paradata to determine the best time for our enumerators to contact households. This information is then combined with household data,

[4] The American Community Survey or ACS is a nationwide survey that collects and produces information on demographic, social, economic, and housing characteristics about our nation's population every year. The Census Bureau contacts about 3.5 million households each year to participate in the survey. The ACS replaced the census long form starting with the 2010 Census.

enumerator data, and business stopping rules to determine an enumerator’s case assignments and the route they should follow each day.

Inputs into the model include, GIS inputs, 2010 Census data (Census response data, payroll data, employee location data), information from the Master Address File, and paradata from the American Community Survey. The resulting outputs of this model are an assigned daily case list for enumerators, predicted travel time for all legs within the optimized route, estimated time to find location for all case contacts and estimated interview time and contact outcome for all case contacts (in order to set standard for number of cases to provide enumerators per shift). We have conducted one successful large-scale simulation of this prototype and are putting it in the field for the first time later next month in the 2015 Census Test. We are very excited about the promise our new control system holds for the Census and the surveys conducted by the Census Bureau. The test this year is very important, but it is just the first test in the field for this prototype system.

I could continue with additional examples of our reengineering efforts and the integration of big data into the planning of the 2020 Census, but I think I will pause here. As I’ve discussed, for our 2020 Census we are focusing our research and testing in a number of areas. We believe that the availability of technology, combined with the data that are now available, will allow for a large number of enhancements to our processes and to the 2020 Census.

Thank you.