

2020 Field Reengineering – ROCKIT

Stephanie Studds
Business Team Lead
U.S. Census Bureau

Presented to the Census Scientific Advisory Committee
September 18th, 2014

The constitutional requirement to conduct a decennial census is becoming more complex and costly due to an increasing diversity in our population. The U.S. Census Bureau (“Census Bureau”) was mandated to perform the 2020 Decennial Census at the same cost, adjusted for inflation, as realized in the 2010 Decennial Census. This constrained budget, coupled with the difficulty in accurately and efficiently counting a diverse population, has led the Census Bureau to rethink its approach for how it conducts certain components of its field operations.

The Census Bureau is addressing this challenge through the 2020 Field Reengineering initiative. This initiative aims to realize process and cost efficiencies within field operations, specifically in the non-response follow up area. The effort evaluates the feasibility of fully utilizing the advantages of planned automation and available real-time data to transform the efficiency and effectiveness of data collection operations.

Discussion Questions:

1. What is the best source for real time data related to traffic? Is it worth doing it? What we mean by this is in a 6 week mission, which has daily case attempts for the enumerators, will we gain enough efficiencies and benefits to outweigh the risks and time it would take to integrate. We believe this would potentially be a small segment of the larger picture. Will this really have a significant impact?
2. So currently we are pushing work assignments (workload - more than can be worked in a shift) from MOJO, the operational control system (OCS) to COMPASS, the collection application hosted on a hand held device, once a day. COMPASS is pushing data back to MOJO at a minimum of every 20 minutes. MOJO then displays operational related data for real time management decision making. So looking ahead to the future, should we push to COMPASS more than once per day.
 - a. So the concern here is with connectivity is it worth pushing work assignments one assignment at a time to each enumerator?
 - b. Should we push a full workload and then begin removing assignments as receipts are received from other modes - in real time? Concern here is an enumerator could be in route to an assignment when we could be removing or altering their assignment.
 - c. We do paired interviews, two or more enumerators working together to complete all assignments in a shared facility. This would be because we may only have a few hours or a day window to complete collection for all units in a large scale complex. Should we assign the same workload to multiple enumerators and then make real time adjustments to the assignments as enumerators’ complete work?

Census Scientific Advisory Committee
2014 Fall Meeting: Synopses

3. Given the automation we are doing with payroll, scheduling, operational control system, and the hand-held device we will get the majority of the savings we are looking for. That being said, the remaining delta of savings is related to the real time traffic and push of assignment updates and removals as detailed above. Is the complexity required to implement for real time traffic and push of assignment updates and removals, worth the small percentage gain for a one time blitz mission (6 week operational window), where 300,000 to 600,000 enumerators are hired?

2015 Address Validation Test (AVT)

Patrick J. Cantwell
Decennial Statistical Studies Division
U.S. Census Bureau

Michael R. Ratcliffe
Geocartographic Products and Criteria, Geography Division
U.S. Census Bureau

Presented to the Census Scientific Advisory Committee
September 18, 2014

In the fall of 2014, the U.S. Census Bureau will conduct the 2015 Address Validation Test (AVT) in two parts to help us plan the Address Canvassing operation for the 2020 Census.

In the first part, Census Bureau listers will canvass 10,100 blocks selected across the U.S. (excluding Alaska, Hawaii, and Puerto Rico) via a stratified, systematic sample. The listers will locate, update, add, or delete addresses currently on the Master Address File (MAF). By comparing field results to results predicted by statistical models, this activity will help us assess our ability to use these statistical models (1) to measure error in the MAF during the decade, and (2) to identify areas to be canvassed in the 2020 Address Canvassing operation. The models studied to date use available auxiliary data, and are mainly of two types: logistic regression at the block level, modeling the propensity of an error (e.g., missing five or more valid addresses from the MAF in a given block); and distributional, modeling the probabilities of 0, 1, 2, ..., errors (e.g., via a zero-inflated negative binomial distribution).

In the second part of the test, referred to as Partial Block Canvassing (PBC), Census Bureau geographers or other staff will list portions of blocks. Partial block canvassing focuses fieldwork on a specified location or area within a census block as opposed to traditional address canvassing, in which the field worker traverses the entire census block. PBC depends on a preceding in-office operation that compares the number of addresses contained in the MAF for any given block with numbers of housing units visible in imagery and identifies areas and locations to be focused upon. PBC, therefore, is part of a more comprehensive process for detecting change in the residential landscape and analyzing quality and completeness of the address list. The purpose of this part is to assess (1) our use of imagery and other geographical information to identify areas in which the number of housing units has remained stable as

Census Scientific Advisory Committee
2014 Fall Meeting: Synopses

well as areas experiencing change not reflected in the MAF, and (2) our ability to accurately identify and canvass only the portions of blocks in which changes are concentrated. For efficiency and to obtain more accurate information, some of these blocks will be the same as those in the first part.

Discussion Questions:

1. Our original statistical models used auxiliary data (independent variables, x) available before the Address Canvassing operation conducted in 2009 as part of the 2010 Census, in an effort to predict the errors (dependent variables, y) observed in the 2009 operation. More current models use (we hope) more predictive auxiliary variables, but we won't have the results until after the Address Validation Test, at which time we can develop better models with data obtained from the test. As the models evolve over the decade with new auxiliary data, how can we properly assess how well they will predict errors in the MAF as we prepare for the Address Canvassing operation in 2019, that is, before we know the truth on the ground?
2. A goal for both our statistical modeling and geographical efforts is to predict, or anticipate, where changes to the residential landscape might occur. Can you suggest how we might use economic data, such as changes in land values or other data measuring economic pressures leading to development or redevelopment, to predict change?
3. In identifying blocks for the Partial Block Canvassing Test, we have focused on blocks in which changes are clustered in one portion of the block, thus avoiding the cost of traversing the entirety of the block. Should we also include blocks in which changes are clustered in multiple portions? In such blocks, the canvasser might still traverse the entirety of the block, but would not spend time listing each unit.

2015 Administrative Record Modeling

Vincent Thomas Mule, Jr.
Team Leader, Administrative Records Modeling Team
U.S. Census Bureau

Presented to the Census Scientific Advisory Committee
September 18, 2014

Thomas Mule will provide updates on the research plans using administrative records for the 2015 Census Test, scheduled to begin in April 2014. This presentation will provide an overview of the different ways that administrative records are being utilized during this test. The discussion will include using administrative records to reduce contacts, identify best contact times for interviewing and compensate for unresolved contacts and characteristics.

Census Scientific Advisory Committee
2014 Fall Meeting: Synopses

Discussion Questions:

1. What is your reaction to how we could possibly use administrative records to eliminate or reduce the number of contacts during NRFU?
 2. What is your reaction to how we are using administrative records to help determine the best time to contact units during NRFU?
 3. What is your reaction to how we could possibly use administrative records for unresolved units after NRFU is completed or for missing characteristics?
-

2015 Optimizing Self-Response Test (Non-ID Processing)

Jennifer W. Reichert
Assistant Division Chief, Decennial Management Division
U.S. Census Bureau

Frank McPhillips
Decennial Management Division
U.S. Census Bureau

Presented to the Census Scientific Advisory Committee
September 18, 2014

Jennifer Reichert and Frank McPhillips, Decennial Management Division, will provide updates on the plans for the 2015 Optimizing Self-Response Census Test, scheduled to begin in February 2015. This presentation will provide an overview of the test, focusing on the testing of real-time processing for internet responses without a pre-assigned Census identification number.

Discussion Questions:

1. What messaging could be effective in encouraging respondents to pre-register their email or telephone number to allow the Census Bureau to use electronic notification methods?
 2. What are effective methods for validating responses received without a pre-assigned Census identification number? How do we know that the person responding is who they say they are and that they really live at the address they provide?
-

BIG Data

Ron S. Jarmin
Assistant Director for Research and Methodology
U.S. Census Bureau

Talk less about “Big Data” and more about using technology to modernize how Census collects, processes and disseminates data to improve economic and social measurement.

Need to modernize

- o Costs
 - § Both to reduce costs, but also to
 - § Free up resources to expand and improve the data we provide to our users
- o Declining response rates
- o Competitive challenge from alternative sources of information
- o Possibly co-equal opportunities
- o Our users demand new data products and that data be more timely, available for smaller domains, and incorporate new or substantially modernized items.

How to modernize?

- o Technology offers many opportunities – sometimes referred to as Big Data
- o Many new sources and means of capturing raw source data with which to construct economic and social statistics:
 - § New sources of administrative data (e.g., real estate records)
 - § Social media
 - § Sensor data (e.g., traffic patterns, commodity flows)
 - § Passive collections (e.g., APIs from large firms and organizations, QuickBooks from smaller ones)
- o These newer sources free us from the constraints of paper survey forms
 - § But unstructured (or at least less structured) data pose challenges, including the challenge of developing at least an approximate sampling frame that will support weighting adjustments or melding with traditional survey data.
- o Move from admin data that supports surveys to surveys that support admin/non-survey data

To successfully employ these new opportunities to modernize, we need to make progress on the following distinct but inter-related components:

- o Methodological - how to produce scientifically valid estimates and uncertainty measures of economic and social statistics from data collected from a wide variety of sources, most of which were not designed to produce inputs to the production of official statistics. Empirical and methodological research on bias, variance, and total survey error, will be necessary to support this requirement.
- o Computational - how to develop the hardware and software infrastructure to compute and disseminate statistics constructed from a variety of sources including surveys, administrative sources, transaction data, social media, sensors, and so on.
- o Policy - how to secure legal permissions and stakeholder buy-in to utilize non-traditional sources of data for the production of official statistics. Requirements include legal agreements with data

Census Scientific Advisory Committee
2014 Fall Meeting: Synopses

providers, and engaging the complete set of stakeholders in the legal and privacy space in a transparent way to ensure all understand the cost, benefits and risks of expanding the capabilities of this next-generation federal statistical system.

- o Outreach and marketing - the data products produced employing new data sources and techniques will differ from traditional survey-based statistics. Users will need to be satisfied that these new statistics actually accurately measure the phenomena we intend them to and may in some cases users may need to be educated in how to properly draw inferences from estimates constructed in novel ways.

What are we doing to make progress on these issues and modernize?

- o External collaborations are key
 - § NCRN
 - § RDCs
 - § Other academic collaborations – e.g., MIT, Georgetown, VT, Stanford, Chicago, AIR
 - § Private sector – e.g., Google, ESRI, UPS, First Data
 - § Other agencies – FRB, NIST, OSTP, USPTO
- o Training existing staff
 - § Census / U of Chicago Big Data Class
- o Recruiting new staff with the right skill sets
- o CEDCap and CEDSi
- o Projects –
 - § 2020
 - Admin records
 - Field reengineering
 - § BDS/Patent data linkages
 - § New pilot project with AIR and Big 10 universities to measure downstream impact of federally funded research with “fat pipe” of information on grants.

Expected Outcomes

- o Enhanced understanding of the trade-offs in using all data
 - o Identification of important enhancements to data that will improve quality, for example enhanced meta and para data for non-traditional sources
 - o Changed business processes and staff development
 - o Modernization of hardware, software and personware
 - o Maintenance or enhancement of the role of federal statistical agencies and their data products
-