Nonresponse Followup Modeling and Microsimulation: Examining Cost-Benefit Tradeoffs for 2020

Kevin M. Shaw & John L. Boies
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Abstract

The 2010 Nonresponse Followup (NRFU) operation, at $1.6 billion in execution costs, was the single most expensive operation in the 2010 Census and thus a primary target for cost avoidance research. The use of statistical modeling to predict the likelihood of whether a Housing Unit (HU) in the NRFU operation is vacant or nonexistent, and whether data from Administrative Records (AR) for an individual HU are accurate offers substantial opportunities for examining the cost-benefit tradeoffs of alternative operational designs. This paper describes HU status and AR concordance modeling and the resulting outcomes from microsimulations conducted on the 2010 Census. Following the general strategy the Census Bureau used in the Targeted Address Canvassing (TAC) research, we use data from a range of government AR and census sources as predictor variables in binomial logistic regression models to predict 2010 NRFU HU status outcomes (vacant, deleted, occupied) and 2010 NRFU household size agreement with AR data. The universe for this research is the 47,197,405 HUs identified as the NRFU analytic universe for the 2010 NRFU Operational Assessment. Population count, census Mailout/Mailback outcomes, and address quality indicators at the HU level were all found to be important predictors of NRFU HU status. HU-level cost data from numerous sources were then used to estimate the potential cost avoidance. This paper examines cost-benefit tradeoffs – coverage degradation vs. dollar cost avoidance – based on preliminary models to predict HU status and whether and where to use AR data for NRFU.

Keywords: Decennial Census, Census 2010, 2020 Census, Administrative Records, StARS, Nonresponse Followup, NRFU, Operational Research, Statistical Modeling, AR Concordance Modeling, Cost Benefit Analysis, AR Enumeration, Simulation, Microsimulation, Net Coverage, Statistical Workload Reduction, Cost Avoidance, Targeted Address Canvassing, TAC

1. Motivation

In less than one year after the U.S. Census Bureau delivered the 2010 Census apportionment counts to President Barack Obama on December 21, 2010, the agency began a Research and Testing (R&T) phase to support the 2020 Decennial Census. The R&T plans include researching new listing and enumeration methodologies, and developing more cost-effective Information Technology (IT) systems to sustain those and other census design changes. Seeing results from the R&T phase, we believe some of the more substantive design revisions will arrive by leveraging the use of Administrative

Records (AR) data. Some of this strategic application of AR data can already be seen in the agency's Targeted Address Canvassing (TAC) research (Boies, Shaw, Holland, 2012), where 2009 microsimulations using statistical models predicting where address canvassing should and should have *not* occurred indicated cost-benefit tradeoffs of between $117 and $250 million (in 2009 dollars) with a 0.47 percent increase in the Housing Unit (HU) gross undercoverage rate. The research documented in this paper demonstrates how similar predictive data modeling can be used to avoid costs in the Nonresponse Followup (NRFU) operation, by using AR data to predict HU status (vacant, delete, occupied) and AR household size agreement (with Census results) – referred to as AR concordance modeling. These types of research efforts – simulating how new data sources can be strategically applied to census operations, and measuring the cost-benefit outcomes (coverage and data quality degradation vs. dollar cost avoidance) – will feed the data-driven decisions to meet the cost control goals for the 2020 Census.

Why is there such a large focus on cost avoidance?

Historically, many of the statistical debates at the agency have centered on accuracy. Bias, adjustment and differential undercount are common census concerns that have, and will continue to drive design decisions. It is not very remarkable to note how many of these issues can be condensed into determining the best method for "counting everyone once and only once." But, the latter half of that equation – "at what cost?" – has now become an essential ingredient; especially in these economic times when many federal agencies are having to do more with fewer resources. In July 2012, in his last Congressional testimony, former Census Bureau Director Dr. Robert Groves said: "For the first time, this decade, we'll have cost-quality tradeoff (information by operation) … We've never had the discussion in this country: How good does the census need to be at a particular cost? … We'll need Congressional help on this … It's a tough tradeoff decision, and it belongs to Congress I believe." (Groves 2012). Integrating concepts like cost-benefit, cost-effectiveness, and return on investment directly into the statistical design, discussion and results is an exciting prospect. One can easily imagine how a menu of census design options that clearly connects the data quality (coverage, etc.) and dollar cost of each menu item, will supply senior management within the agency and department, our nation's lawmakers and other census stakeholders the information they need to make the set of decisions that chart the course for the next census.

Figure 1. Rising Census Costs 1940-2010

Census	Total Population	Total Housing Units	Census Cost (in millions)		Cost per Person	Avg Cost per Person	Cost per HU	Avg Cost per HU
			Unadjusted	In 2010 dollars				
1940	131,669,275	37,438,714	67.5	1,051.5	$ 7.98	$ 6.39	$ 28.05	$ 20.60
1950	151,325,798	46,137,076	91.5	844.9	$ 5.48		$ 17.96	
1960	179,323,175	58,326,357	127.9	945.4	$ 5.25		$ 16.15	
1970	203,302,031	68,704,315	247.7	1,402.4	$ 6.84		$ 20.25	
1980	226,542,199	88,410,627	1,078.5	2,902.7	$ 12.60	$ 12.60	$ 32.29	$ 32.29
1990	248,718,301	102,263,678	2,498.8	4,216.1	$ 16.73	$ 16.73	$ 40.68	$ 40.68
2000*	285,230,516	117,323,117	6,377.8	8,035.1	$ 28.17	$ 28.17	$ 68.49	$ 68.49
2010*	312,471,327	133,341,676	12,382.1	12,594.4	$ 40.31	$ 40.31	$ 94.45	$ 94.45

*Census 2000 and 2010 population, housing and cost figures include Puerto Rico. Census 2010 includes American Community Survey costs.
Source: 1940-1990 census costs from U.S. Census Bureau, "Measuring America: The Decennial Censuses: 1790-2000," Issued September 2002; inflation calculated using BLS Data Series CUUR0000AA0 - CPI (All Urban Consumers), 1967=100. 2000 and 2010 census unadjusted and adjusted costs from U.S. Census Bureau Decennial Management Division (DMD), May 2013; inflation calculated using Office of Management and Budget (OMB) Chained CPI.

Given the soaring per unit census costs (both per household and per person) observed in the last four decennial censuses, it is not surprising that a great deal of emphasis is, and will continue to be, on new processes and techniques to reduce costs while maintaining an acceptable level of accuracy. Substantially decelerating and ideally halting this trend in rising census costs is one of the primary drivers for the R&T work. From the inflation-adjusted values in Figure 1, compared to the $20.60 average per HU cost of the census 1940 through 1970, the per HU census cost increased by a factor of 1.6 in 1980, 2.0 in 1990, 3.3 in 2000 and 4.6 in 2010. In 2010, totaling approximately $12.6 billion, the census cost nearly $100 per HU. Compared to 1970, after adjusting for inflation, this is about nine times as costly in total and nearly five times as costly per HU. With these sharp increases, many are interested to understand census cost drivers, and develop less-costly techniques for creating the product at the cornerstone of our democracy that determines apportionment, redistricting and federals funds allocations.

What census operations are good targets for cost avoidance research?

In the 2010 Census, the two most costly field activities were the Address Canvassing (AC) and NRFU operations. Estimating indirect costs of each operation at a multiple of 0.84 direct costs (Holland 2012), in combination these two operations totaled approximately $4.5 billion – over 36 percent of the total cost of the 2010 Census. It is for this reason that these operations have received priority for conducting operational research to support census design changes in 2020. And, since listing and nonresponse activities are also very costly for current surveys, there is much potential for this research to have measurable and substantial utility in the intercensal period as well.

2. Background

The NRFU operation is born out of households not participating in the self-response phase of the Census. For most in 2010, this self-response phase was the Mailout/Mailback (MO/MB) operation; where households were mailed the short 10-question 2010 Census form, and were instructed to return it in a postage-paid envelope. When households did not return their 2010 questionnaire or complete the form via telephone, NRFU enumerators were instructed to collect the census information on paper forms through personal visits, telephone calls or proxy respondents (a knowledgeable source such as a neighbor). The final 2010 Census mail response rate was 66.5 percent (86.5/130.0 million HUs), and the final mail return rate that excludes about 21.2 million vacant, Undeliverable As Addressed (UAA) and Update/Leave (U/L) and Urban Update Leave (UU/L) deleted HUs was approximately 79.4 percent (86.4/108.9 million HUs) (Letourneau 2012). From the entire MO/MB universe, about one-third did not return their questionnaire. In 2010, the NRFU universe was 47,197,405 HUs, which includes about 3.9 million HUs that returned their form after April 19th. The final Census 2000 response and return rates were very comparable, at 67.4 and 78.4 percent respectively (with corresponding short form rates at 69.1 and 80.1 percent) (Treat 2004).

From the 2010 Nonresponse Followup Operations (NRO) Assessment Report (Walker, et al., 2012), we have numerous metrics of the NRFU operation – The operation began one month after Census Day on May 1, 2010; with all but one of the 494 Local Census Offices (LCOs) completing the operation about two months later on July 9th. Enumerators made up to six contact attempts, where the first attempt was required to be an in-person visit. Excluding the secondary NRFU operations, there were over 515,000 enumerators

and nearly 100,000 supervisory staff in the field (crew leaders, crew leader assistants and field operation supervisors) that worked over 16.6 million training hours and 68.7 million production hours (in total, almost 9,740 years measured sequentially), and drove over 363.6 million miles. In perspective, the number of employees hired to conduct the 2010 NRFU operation is nearly the same quantity of persons living in some of our least populous states; such as Wyoming, District of Columbia, North Dakota and Alaska. And, the amount of miles driven is the equivalent to driving from New York City to Los Angeles, CA over 130,000 times, traveling the equatorial circumference of the Earth over 14,000 times, or making about two roundtrips to the Sun. By many measures, this was a very large undertaking.

What was the workload and cost of the NRFU operation?

In 2000 and in 2010, the main NRFU workload (excluding auxiliary NRFU operations) represented about 36 percent of the total number of final Census HUs – about 42/116 million HUs in 2000, and 47/132 million HUs in 2010. From Moul (2002) and Walker, et al. (2012), direct costs for the NRFU operation totaled $1.1 billion in Census 2000 (stateside only, in 2000 dollars), and $1.6 billion in the 2010 Census (stateside and Puerto Rico). Using an inflation adjustment of 1.27 for the period 2000-2010 (BLS CPI data series #CUUR0000AA0), the NRFU direct cost per HU remained relatively flat – approximately $31.35 per HU in 2000 and $33.65 per HU in 2010. In 2000, the direct costs of the main NRFU operation represented about one-sixth of the total census cost, and in 2010 approximately one-eighth. Both of these figures exclude auxiliary NRFU operations – NRFU Vacant Delete Check (VDC) 2010 (a secondary check on HUs identified as vacant or delete, and a first time enumeration of HUs not included in NRFU), NRFU Reinterview 2010 (a quality check on NRFU enumerations), Coverage Improvement Followup (CIFU) 2000, NRFU Residual 2010 and 2000 (additional cases that require re-enumeration, and a first time enumeration of HUs not included in NRFU and VDC), and NRFU POP99 for 2000 (occupied HUs with unknown population). In 2010, direct costs of these secondary NRFU operations totaled approximately $420 million, in addition to the $1.6 billion cost. In 2000, it was another $250 million above the $1.1 billion cost (in 2000 dollars). All of these costs represent direct costs only (also known as the execution costs) – principally, enumerator and field management staff training, salary (including FICA) and mileage reimbursement.

In addition to the direct costs, a sizeable amount of indirect or overhead costs – e.g., materials/equipment, infrastructure and contract costs – associated with each of these operations was also incurred. As referenced earlier, indirect costs have been estimated at a multiple of 0.84 direct costs (Holland 2012). Factoring in both the cost of the auxiliary NRFU operations and estimating the indirect costs, the total census nonresponse costs were about $2.5 billion in 2000 (unadjusted) and $3.7 billion in 2010. As a percentage of the total census costs, NRFU represented 40 percent in 2000, and 30 percent in 2010.

What were the results of the NRFU operation?

Moul (2002) and Walker, et al. (2012) offer a number of NRFU operational results for comparison. In 2010, about 71 million persons were enumerated in NRFU; or about 23 percent of the final census enumeration. In 2000, the result was higher, at slightly under 30 percent of the final 285.2 million person enumeration. In 2010, of the approximately 47 million NRFU HU workload, about 30 percent were identified as vacant and 9 percent

as delete or unresolved. Similarly, in 2000, of the approximately 42 million NRFU HU workload, 37 percent was identified as vacant (23 percent), deleted and unresolved (14 percent). In 2010, about 52 percent of all NRFU HUs and 24 percent of occupied HUs were completed by proxy. In 2000, proxy rates were markedly lower at only 37 percent for all HUs and 16 percent for occupied HUs. It is important to note that nearly all of the vacant and deleted HU outcomes were categorized as proxy respondents. In both 2010 and 2000, enumerators were trained to make up to six contact attempts per HU. In 2010, 41 percent of HUs required one contact attempt (a required in-person first visit), about 25 percent required two contacts, 16 percent three and the remaining 17 percent workload balance required four or more contacts. Of the nearly 105 million contact attempts made in the 2010 NRFU operation, 90 percent were in-person visits.

3. Methods

There is much that can be done to reduce the NRFU workload and costs in 2020. And, numerous options *are currently* under consideration. From the operational results and comparisons above, some of the more apparent design change options include: (a) reducing the number of required contacts down from six, (b) re-ordering the modes (e.g., instead of requiring an in-person visit for the first contact attempt, as was implemented in 2010, require one or more less-expensive telephone or email contacts prior to an in-person visit), (c) implementing a tailored mode approach (e.g., recent adaptive design projects are investigating how demographic and other household characteristics may drive mode selection and ordering; HU by HU, or by some aggregate level of geography), and (d) standing up and staffing one or more telephone centers, whereby operators laboring under a lower hourly wage than traditional NRFU enumerators (where in 2010, the average wage across all cases was $14.68 per hour) may be successful in completing a large percentage of the NRFU workload. These are all sound options, with one or more very likely to be seen in mid-decade tests and even in 2020 production. And, surely there are cost reducing concepts yet to be brainstormed.

What high-level process does the methodology presented here follow?

Here, we present a methodology that has rich opportunities for synergy with the options just mentioned as well as other potential design changes. The method described in this paper pursues the direct use of AR data sources, both as an input into predictive models, and for enumeration (household status, household population size and roster information). The process for this methodology is:

1) **Construct a census-like AR database –**
   Integrate census and one or more AR data sources to construct an unduplicated, composite household file with rosters and demographic characteristics;
2) **Establish dependent and independent variables for model building –**
   Define desired census operational outcomes (from listing or enumeration) and create individual-level (person, household) and aggregate (census block, etc.) predictors, both annually and longitudinally;
3) **Develop scientific models and set a probability threshold –**
   Use the predictors to develop scientific models for determining where the AR data are accurate (e.g., where the AR HU status and population size agree with known Census response data), and establish a threshold for an acceptable predicted probability value from the model; and

4) **Use AR roster data for enumeration and evaluate the outcome –**

   Where the cases fall above the predicted probability threshold, use the AR data (e.g., HU status, household size, roster characteristics) for enumeration purposes, and evaluate the performance by conducting a cost-benefit microsimulation.

We refer to the use of AR data in Step (4) as an ***AR enumeration***, consistent with the constitutionally-mandated 'direct enumeration.' Here, known limitations with AR data accuracy are addressed by predicting where the AR data are sufficient and accurate for census enumeration purposes. HUs with known poor quality AR data (e.g., group quarters, lower-income households) are driven towards other NRFU modes (phone, personal visit, etc.) to collect accurate responses, while HUs with high-quality AR data are enumerated directly with the AR roster information.

To accomplish this, numerous Census and AR data sources were compiled and mined to derive a census-like database, from which an analytic database of dependent and independent variables was created (Steps 1 and 2). The database of predictors in Step 2 formed the foundation from which we developed the two binary logistic regression models (Step 3) presented in this paper. This regression, modeling the answer to the question: "Are the AR data accurate?" results in:

$$Ln\,[\,p\,/\,(1-p)\,] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ \ldots\ b_kx_k + e,$$

where p represents the probability of the event (an affirmative answer to the question posed by the dependent variable), b_0 is the y-intercept, x_1 to x_k represent the independent variables in the model, the coefficients b_1 to b_k are each independent variable's association with the outcome (log odds), and an error term (e) to capture the differences between the observed and model-predicted values.

We refer to the process of predicting where AR data are in agreement with, or within a pre-determined tolerance of, Census response data as ***AR concordance modeling***. These concordance models are used to predict different operational outcomes. The indicators include demographic (e.g., age, race, Hispanic Origin, etc.) and housing characteristics (e.g., urban/rural, structure type, number of units within the structure, etc.); for individuals, households/rosters and at various levels of aggregation (e.g., census block). Only data prior to the Census Day reference date (April 1, 2010) were used. We also performed these identical steps for the TAC operational research (Boies, Shaw, Holland, 2012). For the two operations (NRFU and AC) that represented over 36 percent of the total cost of the 2010 Census, the set of operational outcomes (dependent variables) ranged from the number of added and deleted addresses captured in a census block during listing (for the TAC research), to the correct HU status (occupied, vacant, delete) and correct HU population count (for the NRFU research). Where these data can adequately predict operational outcomes, the strategy is simple – the operational workload can be decreased via AR enumerations, thereby achieving substantial cost avoidance results. In the TAC research, this amounted to using the database of indicators to predict where address canvassing should and *should not* occur. And, for the NRFU research in this paper, the workload reduction arrives by using the database of indicators to predict where the AR HU status and household/roster count are accurate and thus can be used directly for enumeration purposes. Using these ***Statistical Workload Reduction (SWR) models***, we are able to quantify the cost-benefit outcomes of performing less than 100 percent of the work under traditional decennial procedures.

What AR data sources are viable for this operational research and predictive analytics?

When considering what AR data to use for this research, we defined AR data as any electronic information (record) collected to administer a program or support a business (e.g., information contained on tax returns collected to execute tax policy, account information from your phone/internet service provider collected for billing, service management and marketing purposes). With this definition, there are numerous AR data sources to consider (including data from the Census Bureau itself – survey and census data are indeed collected to administer and execute the responsibilities of the agency).

In 1999, a small staff within the Planning, Research and Evaluation Division at the Census Bureau began building a composite database – the Statistical Administrative Records System (StARS). Constructed from eight national datasets, provided for under agreements with six agencies, the StARS database was designed to cover as broad a spectrum of the U.S. population as possible. StARS databases were produced annually, for 1999 through 2010, and contained records from the:

1. Internal Revenue Service (IRS) Individual Master File (1040),
2. IRS Information Returns File (W-2/1099),
3. Center for Medicare and Medicaid Services (CMS)
   Medicare Enrollment Database (MEDB) file,
4. Indian Health Services (IHS) Patient Registration System file,
5. Selective Service System (SSS) Registration file,
6. Department of Housing and Urban Development (HUD)
   Tenant Rental Assistance Certification System (TRACS) file,
7. HUD Multifamily Tenant Characteristics System (MTCS) file, and
8. Social Security Administration (SSA) Numerical Identification (NUMIDENT) file.

The integration of files 1 through 7, most received by the Census Bureau in April of each year, edited and verified against file 8, matched against the internal Census Master Address File (MAF) Topologically Integrated Geographic Encoding and Referencing (TIGER) database known as the MTdb, and unduplicated, would produce two primary StARS products each year: (1) The Composite Person Record (CPR) files – where each entry represents one unique, validated AR person record; and (2) The Master Housing File (MHF) – where each entry represents one unique AR address (housing unit) record.

Given the more than a decade investment in building and documenting these StARS products (1998 through 2011), we began with these data files, together with Census 2010 data files (pre- and post-operation) and the MAF, as a foundation for our research. The definitions of the StARS products closely align with decennial products, and the StARS products have the important quality of being largely created independently from the Census products. From 2000 through 2010, StARS population totals and characteristics track very closely with decennial census and intercensal population estimate results. In part, this confirms the original design goal to optimize coverage.

What statistical models and cost measures were developed?

Numerous predictive models were developed during the course of this research, and two of the top-performing models were selected for presentation here. These two models are both binary logistic regression models, developed using the entire data universe of 47 million HUs. The first model, model #SB001, predicts vacant HUs. This was defined as where the Census 2010 HU count equals zero. The second, model #SB002, predicts an occupied HU with the same Census 2010 HU count as observed in the StARS AR data. This was defined as where the Census 2010 HU count equals the StARS HU count. Descriptive statistics and odds ratios for the variables in these models can be found in Figure 2. The odds ratio provides the increase or decrease in the likelihood, resulting from a unit change in the independent variable, that the event of interest will

Figure 2: Dependant and Independent Variable Summaries for SWR Models #SB001 and #SB002

	Dependent Variables		Mean	Std Dev	Legal Values
	Model #SB001 - occupied (used to predict vacancy)		0.636	0.481	0,1
	Model #SB002 - arok (used to predict an accurate AR HU count)		0.204	0.403	0,1
	Independent Variables		**Mean**	**Odds Ratios**	
				SB001	**SB002**
Census / DSF	gqv_unitstat	Feb 2010 MAF Occupied Unit Status=1	0.97	1.438	1.002
	gqv_whenbuilt	Built Since Census 2000=1	0.71	1.095	0.873
	gqv_completecs	Complete City Style Address in MAF=1	0.94	1.181	0.984
	gqv_description	Has Unusual within Structure Identifier on MAF=1	0.02	0.752	0.775
	maf_dsfyes	On the Feb 2010 DSF in MAF=1	0.67	0.838	1.324
	maf_dsfunstable	Not Continuously on the last 4 DSFs from Feb 2010=1	0.06	0.814	1.503
	maf_multi	Mult-Unit Structure on Feb 2010 MAF=1	0.31	1.321	1.212
	maf_excfromstat	Excluded from Delivery Statistics in Feb 2010 DSF=1	0.06	0.569	0.918
	gqv_acdelete	2010 Census Addess Canvassing (AC) Delete=1	0.00	0.946	1.268
	gqv_add	2010 Census AC True Add=1	0.09	0.762	1.261
	gqv_change	2010 Census AC Change=1	0.17	0.892	0.943
	gqv_luca2010	MAF Source is 2010 LUCA=1	0.03	0.896	1.113
	enum_isvacant	MAF variable isvacant Yes=1	0.07	0.376	0.907
	gqv_missdsfrt	Missing value on MAF variable DFSRT=1	0.24	0.91	0.994
	uaa_vacants	UAA Reason Vacant=1	0.20	0.112	0.233
	uaa_deletes	UAA Reasons Correlated with Census Deletes=1	0.15	0.6	0.621
StARS	starspeople	HU person count from StARS, 0-99	1.22	1.218	0.83
	stars_white	StARS Person 1-10 in HU White=1	0.39	1.536	2.683
	goodarok	HU present in StARS 1999-2010=Yes	0.30	0.942	1.079
	aroknotocc	HU never in StARS 1999-2010=1	0.24	0.562	<0.001
	pop4	Person Count the Same for 2007-2010 in StARS=1	0.21	0.81	1.071
	sroster2	Roster the Same for 2007-2010 in StARS=1	0.12	1.95	2.927
	ipres1	Infant Present in StARS HU 2010=1	0.08	0.942	1.369
	rpres2	>65 Present in StARS HU 2010=1	0.02	1.332	1.08
	stars_sss	StARS Source includes Selective Service=1	0.07	0.896	0.45
	stars_med	StARS Source includes Medicare=1	0.11	0.658	0.533
	stars_numsource	Number of StARS Sources, 0-11	1.27	1.147	1.673
	fbpres	Foreign Born Present in StARS HU 2010=1	0.09	1.091	0.846
	sex	Female Person 1 in 2010 StARS=1	0.18	1.094	0.98
	age25	Person 1 in 2010 StARS, 25-50 Years=1	0.26	1.583	6.214
	age50	Person 1 in 2010 StARS, 50-65 Years=1	0.09	1.801	6.569
	age65	Person 1 in 2010 StARS>65 Years=1	0.05	1.402	8.286

Source: Logistic regression modeling output.

occur. Both models used the same set of 32 independent variables; and yielded strong model fits, with max-rescaled R-square values of 0.45 and 0.52 respectively.

Twelve HU-level Cost Measures (CMs) were developed to conduct the cost-benefit analyses for these and other SWR models. The variation observed in these 12 CMs, introduced by the varying degrees of granularity used to produce them, permitted us to examine the cost avoidance potential of our models under numerous assumptions. The 12 CMs consist of six direct CMs (primarily field staff production training, salaries and mileage) and six total CMs (in addition to direct costs, these costs also include materials, equipment, contract and other overhead costs). These CMs were computed at the HU level, where each individual HU in the 2010 NRFU universe received 12 different cost estimates. The control total for each of the six direct CMs across the about 47.2 million NRFU HU universe was $1,589,397,885, and the control for each of the six total CMs was $2,926,653,330 (direct costs * 1.84). In Holland (2012a), the indirect costs for the next single most expensive 2010 Census operation – AC, was measured at a multiple of 0.84 direct costs. While the 2010 Census AC and NRFU operations are different operations, since the 2010 Census NRFU indirect costs have not yet been precisely measured this offers a reasonable approximation for research purposes.

The six direct CMs, also known as the execution costs of the operation, are defined as:

- **EXE01 - National Average:** For this CM, each HU received the same cost of $33.68/HU ($1,589,397,885 direct costs / 47,197,405 HUs);
- **EXE02 - National Average based on number of contact attempts:** Each HU with the same quantity of contact attempts received the same cost, where each contact attempt was valued at $15.22 ($1,589,397,885 direct costs / 104,432,553 total contact attempts);
- **EXE03 - Local Census Office (LCO) Average:** The same calculation as EXE01 is made within each of the 494 LCOs. The 494 LCO direct cost controls were obtained from the 2010 Decennial Applicant Payroll & Processing System (DAPPS) data file. The DAPPS LCO cost controls are the actual expenditures for each LCO;
- **EXE04 - LCO Average based on number of contact attempts:** The same calculation as EXE02 is made within each of the 494 LCOs. The LCO contact attempt controls were calculated using the 2010 Decennial Response File (DRF). The DRF contact and response data are the primary input into determining the final enumeration for each HU;
- **EXE05 - LCO Average based on number of contact attempts (mode adjusted):** This is identical to EXE04, but includes a mode adjustment that weights in-person contacts at a multiple of 1.86 that of telephone contacts;
- **EXE06 - Enumerator IDentification (EID) daily match to DAPPS data:** This CM was created at the NRFU enumerator level, linking each enumerator's daily timesheet to their daily HU contact attempts. This linkage was attempted for all enumerators, across all days of the operation. For each enumerator, for each day, the sum of salary and overhead costs were equally allocated to all contact attempts, and mileage costs were wholly assigned to in-person contact attempts.

These six direct CMs represent a new internal data product (Holland & Shaw, 2012b). For each of these direct CMs, a corresponding total CM (TOT0{X}, where X=1-6) was

created. A simple linear assignment of indirect costs per HU was generated. Where cost avoidance estimates are provided, this results in a corresponding linear reduction in total costs (e.g., where workload is reduced by X percent, it is assumed total costs are reduced at the same rate).

4. Microsimulation

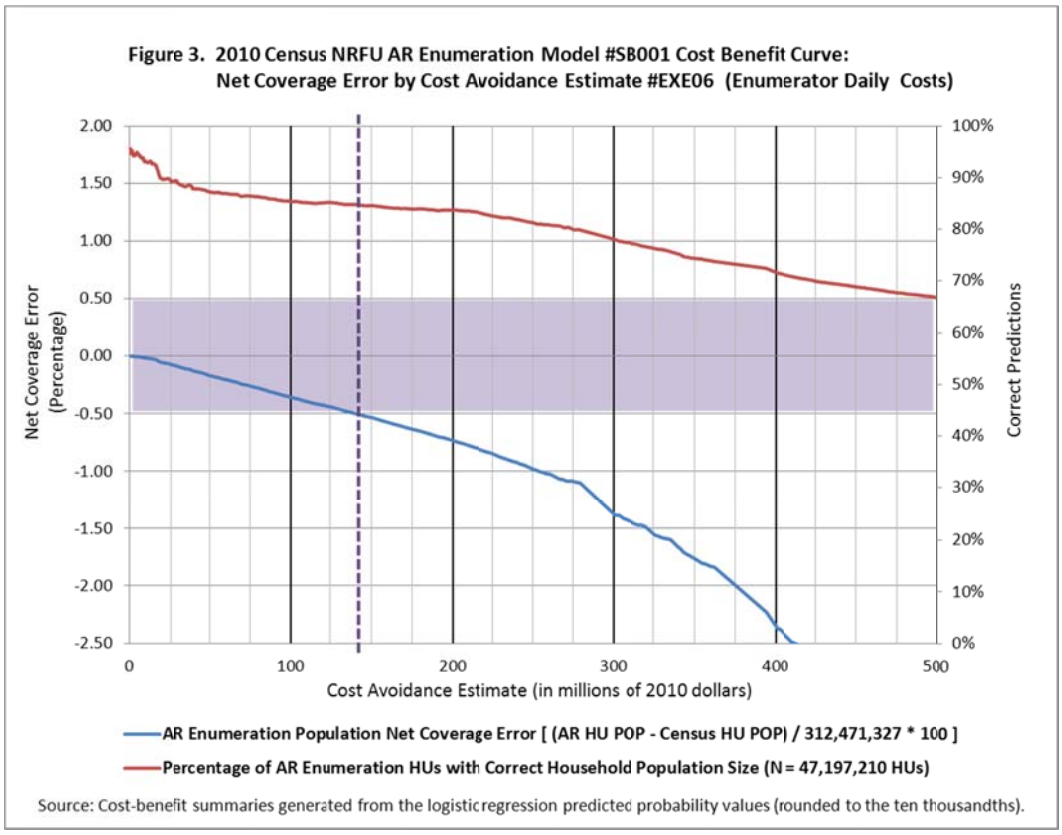
Once the logistic regression models and cost measures were developed, we evaluated the cost/benefit performance of the models. The primary evaluative tool was a microsimulation to conduct the cost-benefit analysis. Here, the microsimulation answers the primary question: "What cost and quality outcomes would have been observed if the model was used in the 2010 NRFU operation?" For this cost-benefit analysis, the 'cost' is defined as coverage degradation, and the 'benefit' is defined as dollar cost avoidance. In other words: Based on our SWR models predicting HU status and whether and where to use AR data for NRFU, what is the cost avoidance potential at various levels of coverage loss? Only net population coverage loss is examined here. For the particular set of HUs upon which the model acts, this is defined as the [AR Enumerated HU Population] / [2010 Census HU Population].

For this microsimulation, we concentrated on the target of a 0.5 percent net population coverage loss. In 2010, this equates to a net positive or negative difference of approximately 1,562,356 persons from the final 2010 Census count of 312,471,327 persons. This is not an agency-established target for 2020, but simply provides a reference point to narrow discussions and compare model results. From the cost benefit analyses here, for either model, any level of coverage degradation (from low to high, however one defines those bounds) can be considered and examined.

Figure 3 depicts the cost-benefit curve for the first model, model #SB001 which uses AR and Census data to predict vacant HUs. The cost avoidance estimate is displayed on the x-axis and the population net coverage error is shown on the first vertical axis. The purple band represents the pre-determined target of +/- 0.5 percent population net coverage error, and the dashed vertical line represents where the coverage error traverses the banded threshold. From left to right, the blue line tracks how much of that error is introduced as more AR enumerations are made. So, as we use AR data to enumerate HUs (i.e., the individual HU roster is populated *solely* with the data in the AR files, without adjustment or augmentation), we observe a shallow but steady departure from the zero-error parallel.

At the point we reach the -0.5 percent net coverage error threshold, the cost avoidance is estimated at about 142.1 million dollars (in 2010 dollars, direct costs only). Here, only CM #EXE06 (the most detailed measure) is used to quantify the cost avoidance potential. Using AR data to predict about 5.2 million vacant/non-existent HUs, this represents an approximately 9 percent reduction in operational costs. For comparison, if one doubles the level of acceptable net coverage error to one percent, the cost avoidance is estimated at 255.1 million dollars, or about 16 percent of the operational costs. Since model #SB001 predicts vacant HUs, by definition, no AR data are used to populate the roster since the HU counts are predicated as zero. At these error levels (-0.5 percent and -1.0 percent), the amount of correct predictions (i.e., the quantity of HUs where the HU count is infact zero) are 85 and 81 percent respectively. From the figure, the red line that displays the correct prediction levels confirms that, in volume, the model is successfully

arraying the 47 million HU universe in a manner that assigns a larger predicted probability to those HUs that have correct AR data. From the perspective that this model is not a far departure from the traditional NRFU approach that only uses data collected by NRFU enumerators to populate the roster, this model is conservative. And at that, alone, it still offers sizable cost avoidance potential.



In contrast to the previous model, Figure 4 portrays a dramatic cost-benefit curve for model #SB002. Model #SB002 predicts an occupied HU with the same HU count as observed in the AR data; where there is concordance between the known Census result and the AR roster count. Here, we observe that the level of correct predictions *enters* the graph at about 80 percent. While the coverage error traverses the target purple band twice before its final intersection, we notice that for the set of HUs that achieve the first $400 million (in 2010 dollars, direct costs only) of cost avoidance, the coverage error nearly overlaps the zero-error parallel for the entire span. While the population net coverage error is virtually nonexistent for this set of AR enumerations (about 10.9 million HUs), we see a sharp degradation in the correct prediction level compared to the previous model. At this level of cost avoidance, only 58 percent of the AR-enumerated HUs have the correct HU count.

For model #SB002, the cost avoidance is estimated at about 874.3 million dollars at the point where we reach the -0.5 percent net coverage error threshold. Using AR data to predict and to populate the roster for approximately 24.0 million HUs, this represents an approximately 55 percent reduction in operational costs. While doubling the level of acceptable net coverage error in the previous model offered a near linear increase in the

cost avoidance potential, that is not observed here. It is estimated that increasing the error level from -0.5 to -1.0 percent would only increase the cost avoidance an additional 2.4 percent (57.4 percent in total). Since model #SB002 predicts occupied HUs, this model offers an aggressive and strategic use of AR data. At the terminal point at which the error crosses the target banded threshold, using AR data to enumerate 24 million HUs equates to about half of the 2010 Census NRFU HU workload. This application of SWR models would have substantial outcomes for the cost and quality of the census.

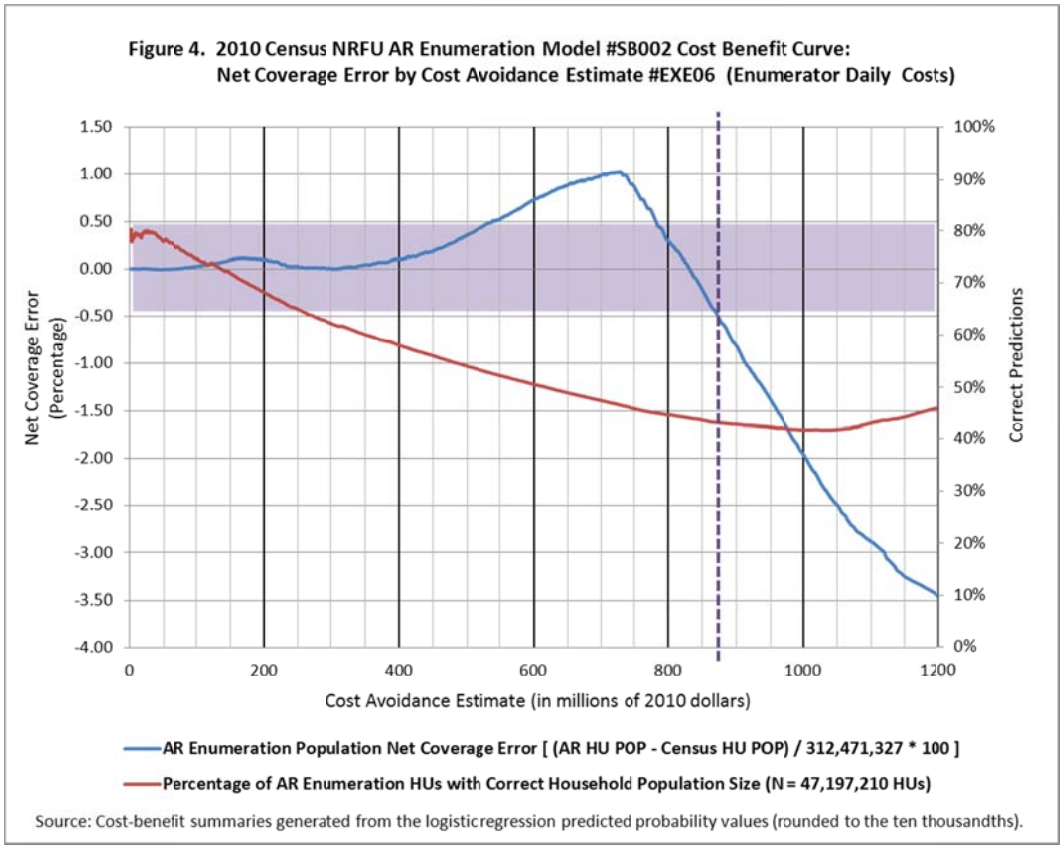


Figure 5 summarizes the results of the microsimulations for both models, and offers various model performance outcomes. By design, there is a stark difference in the universe that each model seeks to make accurate predictions on – zero population HUs in model #SB001 versus non-zero population HUs with accurate AR HU counts in model #SB002. From the figure, both model fits are quite strong, at 0.45 and 0.52 respectively. While the earlier figures allow the exploration of using AR across a wide error spectrum, this table summarizes the performance outcomes at a population net coverage error level of approximately 0.5 percent.

As expected, model #SB001 operates primarily on non-occupied HUs, and thus the resulting distribution of HU status is about 85 percent non-occupied HUs; or 4.5 of the 5.2 million HUs in the applied model universe. This results in over 10.1 million contact attempts avoided. In millions of 2010 dollars, the application of this model at this error level would result in cost avoidance of between 9 and 11 percent of the costs of the operation. This ranges from $142.1 - $175.9 million in direct costs, or $261.6 - $323.9 million in total costs.

Model #SB002 is more aggressive, in that it uses a large amount of AR roster data – a considerable departure from the traditional nonresponse door-to-door data collection. This model operates primarily on occupied HUs, with over 87 percent of the applied model universe occupied in the 2010 Census. However, here the HU-level correct prediction rate drops to 43 percent. Applying model #SB002 would have avoided more than 56.5 million contact attempts – which represents more than 54 percent of all contact attempts made in 2010 Census NRFU operation. At the population net coverage error level of -0.52 percent, the estimated cost avoidance ranges from $806.5 - $874.3 million in direct costs, or $1.5 - $1.6 billion in total costs. This amounts to approximately a 51-55 percent cost avoidance for the operation.

Figure 5: Summary of Simulation Outcomes for SWR Models #SB001 and #SB002

			Logit Model	
	Metric		**SB001**	**SB002**
	Model Fit (Max-Rescaled R-Square)		0.4522	0.5194
	Predicted Probability Threshold		0.1967	0.0526
	Population Net Coverage Error		-0.50%	-0.52%
NRFU Workload Reduction	AR Enumerations (in millions)			
	Housing Units (HUs)		5.2	24.0
	Occupied		0.8	20.8
	Vacant		3.4	2.8
	Delete		1.1	0.3
	Population (persons in HUs)		1.6	58.0
	Percentage of AR Enumeration HUs with Correct Household Population Size		85%	43%
	Contacts Avoided (in millions)		10.1	56.5
	Cost Avoidance Estimate (in millions of 2010 dollars)	EXE01	$ 175.9	$ 806.5
		EXE02	$ 154.3	$ 860.1
		EXE03	$ 167.6	$ 825.0
		EXE04	$ 148.5	$ 866.8
		EXE05	$ 147.9	$ 867.7
		EXE06	$ 142.1	$ 874.3
		TOT1-TOT06	$ 261.6 - $ 323.9	$ 1,485.1 - $ 1,609.9

Source: Mircosimulation results generated from applying the logistic regression models to the study universe.

4.1 Limitations

The research and results presented here should be caveated by the following: (i) None of the implied census design changes have been formally vetted throughout the agency or the wider stakeholder/oversight community, nor have all of the information owners of the various data sources agreed to the referenced *production uses* (this will take considerable time and resources, and likely additional legislation, and should be given priority alongside the technical research leading up to 2020); (ii) Since both of logit models presented here are full population-based models that do not take into account any of the other potential census design changes that may hinder their utility, actual 2020 model performance is unknown; (iii) While the authors place wide utility around the outcomes of the cost estimates used for this research, they were only created to support research decisions related to the viability and *relative performance* of various SWR models;

(iv) Also related to cost estimation, much work remains to estimate indirect operational costs more accurately; and as such the indirect cost estimates should be interpreted with caution, and (v) While we have selected national net population coverage and dollar cost avoidance as the primary cost-benefit metrics, numerous other performance metrics should be considered and quantified in the future (e.g., differential geographic and demographic coverage). See Boies & Shaw (2012b) for a range of examples.

5. Conclusions

For decades, the Census community has praised the *potential use* of AR data in the decennial census. Many have said the use of AR data for enumeration purposes will forever be something for the *next* census. But, here, we present a novel way to overcome some of the known deficiencies in the AR data. The manner in which the SWR models make use of the AR data in this research – using them alongside other covariates to predict where and when the AR data are accurate for a census enumeration – is strategic. In other words, household by household, when we have greater confidence that the AR data are accurate, over other households with AR data, and make the business decision to *only* use the AR data in those instances, we offer a path towards overcoming many of the limitations of the application of AR data in a decennial environment.

In this research, after much approval gathering, data processing, model building and evaluation, we present two final models for consideration. At the population net coverage error level of approximately -0.5 percent, the first model offers cost avoidance potential of about 10 percent, and the second approximately 55 percent. In an operation with direct costs of nearly $1.6 billion, this is substantial. While it is encouraging to record such savings potential at this very low national error level, the HU-level accuracy of between 43 and 85 percent for these models offers much room for improvement. For this approach to have merit in the 2020 Census, these HU-level error rates must be improved; either through model refinement or a supplemental operation. In addition, there is much work to do regarding multi-stage modeling. Through the process to simulate the staging of numerous SWR models into a single 'procedural model,' we can learn how these and other models can be combined, and begin to quantify their synergistic effects. Leading up to the 2020 Census, it will be extremely important to have multiple avenues (even if small-scale) to evaluate and confirm the outcomes of this type of research against traditionally-collected nonresponse data. The Census Bureau already has plans for a site test in 2014, in which these and other models and design changes can be tested.

Essentially, the research performed here created an environment – a methodological process, along with a set of data files – to direct a universe of HUs to a new mode termed *AR Enumeration*, and measure the quality of the HUs enumerated therein. The resulting Census Simulation (CenSIM) database contains integrated, longitudinal intelligence that has much application yet to be explored. Recent discussions have included researching the effects of other and new data files and variables, cost optimization modeling, and establishing one or more Return On Investment (ROI) measures that can accumulate many factors (e.g., net population error, demographic error in aggregate and at the micro-level) and weigh them against the avoided costs of the traditional door-to-door collection. Whatever the priorities and outcomes, the revived potential for AR data to make long-lasting, valuable contributions to the decennial census is exciting!

Acknowledgements

The authors wish to thank numerous Census Bureau employees for their assistance, feedback, encouragement and support; including Jennifer Reichert, Patrick Cantwell, Burton Reist, Frank Vitrano, Claude Jackson and the entire Census Evaluations Branch in the Decennial Statistical Studies Division. Additionally, we extend our gratitude to Jonathan Holland for his thorough and precise work producing the cost measures.

References

Boies, John; Shaw, Kevin M.; Holland, Jonathan (2012a), "2010 Census Address Canvassing Targeting and Cost Reduction Evaluation Report," 2010 Census Program for Evaluations and Experiments (CPEX), #A-03, July 2012.

Boies, John; Shaw, Kevin M. (2012b), "2020 Census 8.107 Supplementing and Supporting Nonresponse with Administrative Records (SSNAR) Simulation Subteam 2010 Census NRFU Model Evaluation Template – Version #1," DSSD Memorandum #DCRT-L2, September 25, 2012.

Holland, Jonathan P. (2012a), "2010 Census Evaluation of Automation in Field Data Collection in Address Canvassing Report," 2010 CPEX Memorandum, July 2012.

Holland, Jonathan; Shaw, Kevin M. (2012b), "2020 Census 8.107 SSNAR Simulation Subteam 2010 Census Nonresponse Followup (NRFU) Unit-Level Cost Methodology & Documentation," DSSD Memorandums #DCRT-L1 & #DCRT-L3: August 2012, May 2013.

GAO-12-80 (2012), "Additional Actions Could Improve the Census Bureau's Ability to Control Costs for the 2020 Census," January 2012.

Groves, Robert (2012), Congressional Testimony: "Census Planning Ahead for 2020," Senate Subcommittee on Federal Financial Management, Government Information, Federal Services, & International Security, July 18, 2012.

Letourneau, Earl (2012), "2010 Census Mail Response/Return Rates Assessment Report," 2010 CPEX Memorandum, May 2012.

Moul, Darlene A. (2002), "Nonresponse Followup for Census 2000, Census 2000 Testing, Experimentation and Evaluation (TXE) Program," July 25, 2002.

Shaw, Kevin M. (2012a), "2020 Census R&T Presentation: 8.107 SSNAR Simulation Subteam Microsimulation Overview and Details," Internal Document, April 2012.

Shaw, Kevin M. (2012b), "Draft 2013 JPSM Practicum Proposal," Internal Document, November 2012.

Shaw, Kevin M. (2012c), "Draft 2020 Census Research and Testing Study Plan: 8.107 SSNAR Simulation Subteam," Internal Document, August 2012.

Treat, James B. (2004). Census 2000 TXE Program Topic Report No. 11, Response Rates and Behavior Analysis, U. S. Census Bureau, Washington, DC 20233, March 2004.

U.S. Census Bureau - Financial and Administrative Systems Division (FASD) (2000), "Management Study of Nonresponse Followup Enumeration," October 2000.

Walker, Shelley, et al. (2012), "2010 Census Nonresponse Followup Operation (NRO) Assessment Report," 2010 CPEX Memorandum, #190, April 23, 2012.

_____. (2013), The Department of Commerce Budget in Brief Fiscal Year 2013. John E. Bryson, Secretary, http://www.census.gov/aboutus/pdf/DOCBudgetinBrief_final.pdf, accessed on August 2, 2013.

*Reports in the 2010 Census Program for Evaluations and Experiments (CPEX) are located at http://www.census.gov/2010census/about/cpex.php.