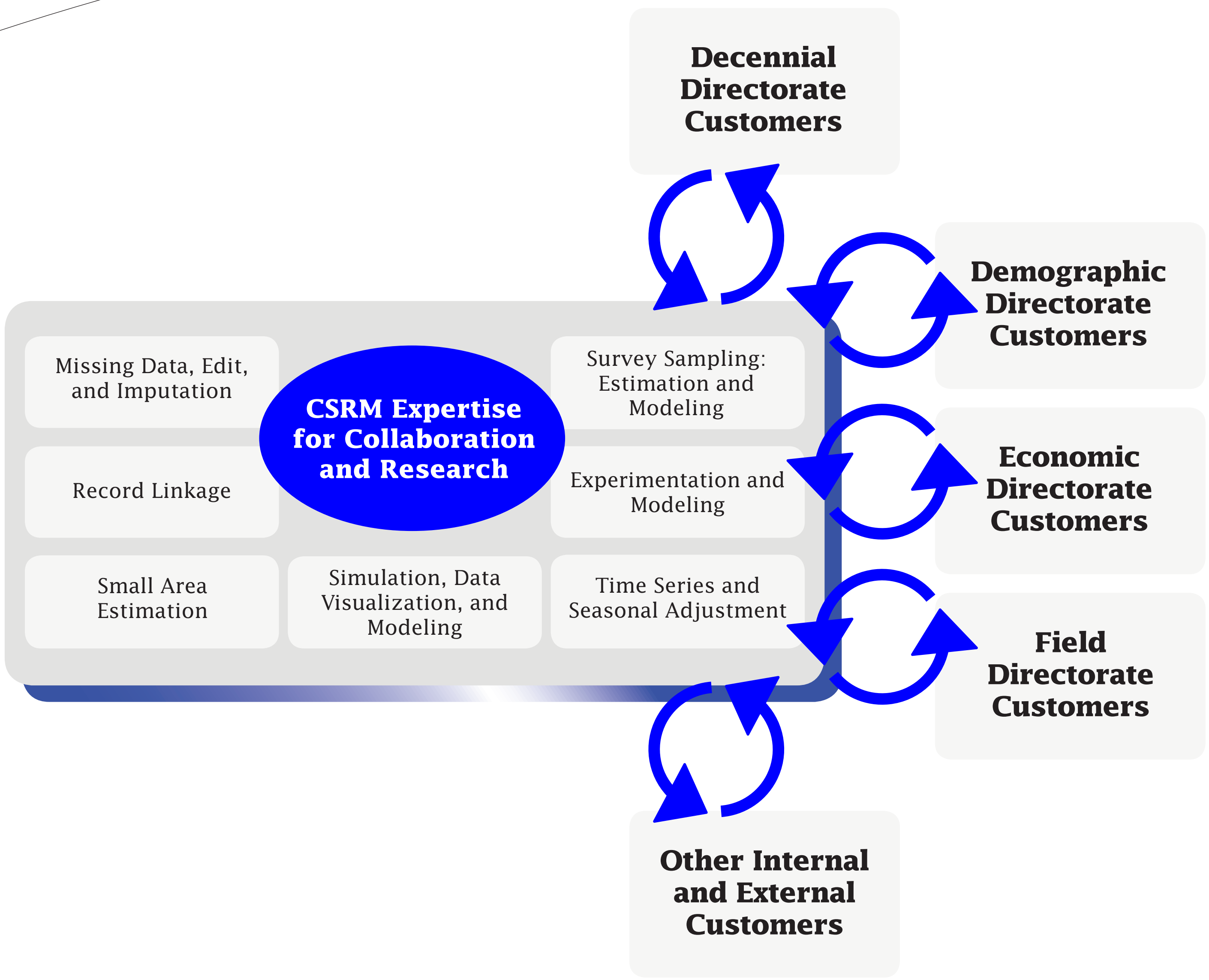
Annual Report *of the* Center for Statistical Research and Methodology

Research and Methodology Directorate

Fiscal Year 2016


Decennial Directorate Customers
Demographic Directorate Customers
Economic Directorate Customers
Field Directorate Customers
Other Internal and External Customers
CSRM Expertise for Collaboration and Research
Missing Data, Edit, and Imputation
Record Linkage
Small Area Estimation
Simulation, Data Visualization, and Modeling
Survey Sampling: Estimation and Modeling
Experimentation and Modeling
Time Series and Seasonal Adjustment

United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Since August 1, 1933—

"... As the major figures from the American Statistical Association (ASA), Social Science Research Council, and new Roosevelt academic advisors discussed the statistical needs of the nation in the spring of 1933, it became clear that the new programs—in particular the National Recovery Administration—would require substantial amounts of data and coordination among statistical programs. Thus in June of 1933, the ASA and the Social Science Research Council officially created the Committee on Government Statistics and Information Services (COGSIS) to serve the statistical needs of the Agriculture, Commerce, Labor, and Interior departments ... COGSIS set ... goals in the field of federal statistics ... (It) wanted new statistical programs—for example, to measure unemployment and address the needs of the unemployed ... (It) wanted a coordinating agency to oversee all statistical programs, and (it) wanted to see statistical research and experimentation organized within the federal government ... In August 1933 Stuart A. Rice, President of the ASA and acting chair of COGSIS, ... (became) assistant director of the (Census) Bureau. Joseph Hill (who had been at the Census Bureau since 1900 and who provided the concepts and early theory for what is now the methodology for apportioning the seats in the U.S. House of Representatives) ... became the head of the new Division of Statistical Research ... Hill could use his considerable expertise to achieve (a) COGSIS goal: the creation of a research arm within the Bureau ..."

Source: Anderson, M. (1988), *The American Census: A Social History*, New Haven: Yale University Press.

Among others and since August 1, 1933, the Statistical Research Division has been a key catalyst for improvements in census taking and sample survey methodology through research at the U.S. Census Bureau. The introduction of major themes for some of this methodological research and development where staff of the Statistical Research Division[1] played significant roles began roughly as noted—

- **Early Years (1933–1960s):** *sampling (measurement of unemployment and 1940 Census); probability sampling theory; nonsampling error research; computing; and data capture.*
- **1960s–1980s:** *self-enumeration; social and behavioral sciences (questionnaire design, measurement error, interviewer selection and training, nonresponse, etc.); undercount measurement, especially at small levels of geography; time series; and seasonal adjustment.*
- **1980s–Early 1990s:** *undercount measurement and adjustment; ethnography; record linkage; and confidentiality and disclosure avoidance.*
- **Mid 1990s–Present:** *small area estimation; missing data and imputation; usability (human-computer interaction); and linguistics, languages, and translations.*

At the beginning of FY 2011, most of the Statistical Research Division became known as the Center for Statistical Research and Methodology. In particular, with the establishment of the Research and Methodology Directorate, the Center for Survey Measurement and the Center for Disclosure Avoidance Research were separated from the Statistical Research Division, and the remaining unit's name became the Center for Statistical Research and Methodology.

[1] The Research Center for Measurement Methods joined the Statistical Research Division in 1980. In addition to a strong interest in sampling and estimation methodology, research largely carried out by mathematical statisticians, the division also has a long tradition of nonsampling error research, largely led by social scientists. Until the late 1970s, research in this domain (e.g., questionnaire design, measurement error, interviewer selection and training, and nonresponse) was carried out in the division's Response Research Staff. Around 1979 this staff split off from the division and became the Center for Human Factors Research. The new center underwent two name changes—first, to the Center for Social Science Research in 1980, and then, in 1983, to the Center for Survey Methods Research before rejoining the division in 1994.

U.S. Census Bureau
Center for Statistical Research and Methodology
Room 5K108
4600 Silver Hill Road
Washington, DC 20233
301-763-1702

We help the Census Bureau improve its processes and products. For fiscal year 2016, this report is an accounting of our work and our results.

Center for Statistical Research & Methodology
www.census.gov/srd/csrm

Highlights of What We Did...

As a technical resource for the Census Bureau, each researcher in our center is asked to do three things: *collaboration/consulting*, *research*, and *professional activities and development*. We serve as members on teams for a variety of projects and/or subprojects.

Highlights of a selected sampling of the many activities and results in which the Center for Statistical Research and Methodology staff members made contributions during FY 2016 follow, and more details are provided within subsequent pages of this report:

- *Missing Data, Edit, and Imputation*: (1) continued research of modeling approaches for using administrative records in lieu of decennial census field visits and documented methodologies in scientific papers; (2) continued investigation of the feasibility of using third party ("big") data (NPD Group, a major credit card, and First Data) to supplement or enhance retail sales estimates in the Monthly/Annual Retail Trade Surveys; and (3) developed a system that generates essentially new implied edits based on given explicit edits.

- *Record Linkage:* (1) applied and made updates to record linkage software; and (2) promoted the use of graphical models with a series of six lectures.

- *Small Area Estimation*: (1) compared small area predictions from a Beta model and a log-linear model for rates; (2) developed a method to estimate design effects of small areas using larger level aggregates to improve design based variance estimation; and (3) developed a Bayesian small area multivariate model which includes measurement error in the covariates.

- *Survey Sampling-Estimation and Modeling*: (1) developed methodology for setting parameters for M-estimation methodology for detecting and treating influential values in an economic sample survey; (2) performed model assessments for alternative statistical models of rates of citizenship, limited English proficiency, and illiteracy within language minority groups in relation to the Census Bureau's mandate to reach determinations on provision of election materials in languages other than English under *Section 203 of the Voting Rights Act*; (3) completed an extensive study of whether mode effects must be accounted for explicitly in survey item imputation within the American Community Survey; and (4) demonstrated that exact optimal sample allocation algorithms all follow explicitly from a simple decomposition of sampling variance.

- *Time Series and Seasonal Adjustment*: (1) published an R user-interface for X-13ARIMA-SEATS, allowing for greater usability and communicability of seasonal adjustments; and (2) developed new estimation for vector time series models, allowing for parameter constraints, enforcement of stability, and co-integration.

- *Experimentation and Statistical Modeling:* (1) formulated a basic decision theoretic framework for block selection with the MAF Error Model; and (2) evaluated spatial models for block-level add counts from 2010 address canvassing.

- *Simulation and Statistical Modeling:* (1) continued developing model based method for analyzing singly imputed synthetic data under multiple linear regression model and multivariate normal models; and (2) evaluated several data visualization methods for statistically comparing populations.

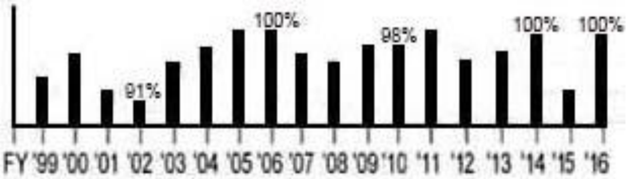
- *SUMMER AT CENSUS:* Sponsored, with divisions around the Census Bureau, scholarly, short-term visits by 37 researchers/leaders who collaborated extensively with us and presented seminars on their research. For a list of the 2016 *SUMMER AT CENSUS* scholars, see *http://www.census.gov/research/summer_at_census/*.

How Did We[1] Do...

For the 18th year, we received feedback from our sponsors. Near the end of fiscal year 2016, our efforts on 30 of our program (Decennial, Demographic, Economic, Administration, External) sponsored projects/subprojects with substantial activity and progress and sponsor feedback (Appendix A) were measured by use of a Project Performance Measurement Questionnaire (Appendix B). Responses to all 30 questionnaires were obtained with the following results (The graph associated with each measure shows the performance measure over the last 18 fiscal years):

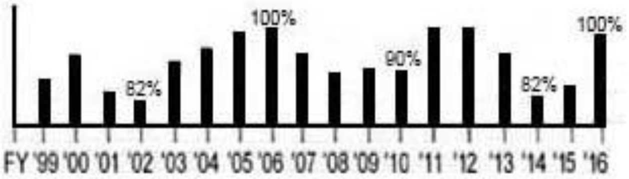
Measure 1. Overall, Work Met Expectations

Percent of FY2016 Program Sponsored Projects/Subprojects where sponsors reported that overall work met their expectations (agree or strongly agree) (30 out of 30) …….…………..... 100%



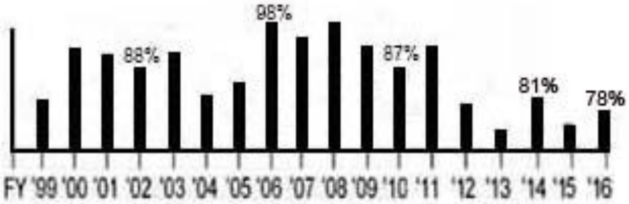
Measure 2. Established Major Deadlines Met

Percent of FY2016 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met (17 out of 17 responses) ……......………..…... 100%



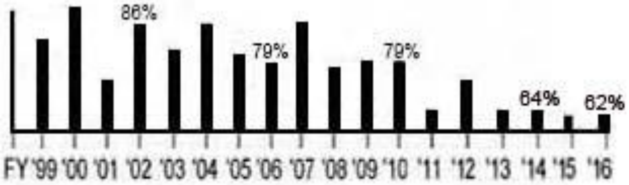
Measure 3a. At Least One Improved Method, Developed Technique , Solution, or New Insight

Percent of FY2016 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight (21 out of 27 responses) ………… 78%



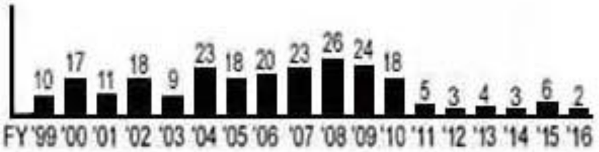
Measure 3b. Plans for Implementation

Of these FY2016 Program Sponsored Projects/Subprojects reporting at least one improved method, technique developed, solution, or new insight, the percent with plans for implementation (13 out of 21 responses) ………………….. 62%



Measure 4. Predict Cost Efficiencies

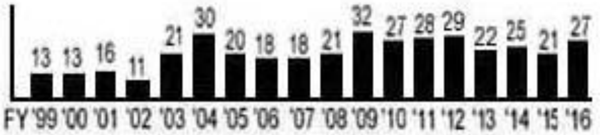
Number of FY2016 Program Sponsored Projects/Subprojects reporting at least one "predicted cost efficiency" ………..…… 2



From Section 3 of this ANNUAL REPORT, we also have:

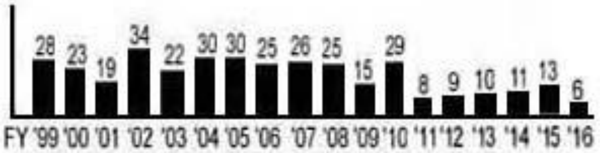
Measure 5. Journal Articles, Publications

Number of peer reviewed journal publications documenting research that appeared (18) or were accepted (9) in FY2016 ……………………………………………………………... 27



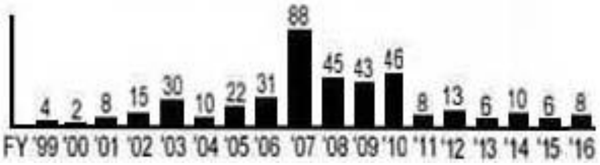
Measure 6. Proceedings, Publications

Number of proceedings publications documenting research that appeared in FY2016 …………………………………..………. 6



Measure 7. Center Research Reports/Studies, Publications

Number of center research reports/studies publications documenting research that appeared in FY2016 …………….. 8



Each completed questionnaire is shared with appropriate staff to help improve our future efforts.

[1] Reorganized from Statistical Research Division to Center for Statistical Research and Methodology, beginning in FY 2011.

TABLE OF CONTENTS

1. COLLABORATION

1.1 REDESIGNING FIELD OPERATIONS (Decennial Project 6650B23)

1.2 ADMINISTRATIVE RECORDS DATA (Decennial Project 6750B01)

1.3 DATA CODING, EDITING, AND IMPUTATION (Decennial Project 6550B01)

1.4 POLICY (Decennial Project 6250B07)

A. Decennial Record Linkage

Description: Under this project, staff will provide advice, develop computer matching systems, and develop and perform analytic methods for adjusting statistical analyses for computer matching error with a decennial focus.

Highlights: During FY 2016, staff provided extensive comments to the Decennial Statistical Studies Division (DSSD) related to background on Decennial record linkage methods and production software. The background covered what name and address standardization software were available and why they were crucial to Decennial processing and research. The background also covered what software (most funded by DSSD) had been written for Decennial processing (the *SRD 1-1 Matcher* used in 1990, 2000, and 2010 and *BigMatch* used in 2010). There are at least seven variants of the software that have been written for DSSD but not used for production and other variants for projects in the Economic Directorate.

Staff wrote a summary related to the accuracy and speed of *BigMatch* that was forwarded to 20+ individuals. In a three-year review by professors at Curtin University in Western Australia, *BigMatch* was considered the most accurate software in comparison with commercial products from IBM and SAS and four shareware products written by university professors for the health agencies. Staff provided advice and details on some of the computational algorithms in *BigMatch*.

Staff wrote a proposal regarding the evaluation and testing of record linkage that was accepted by DSSD. Under the direction of DSSD, staff from several areas (including CSRM) will evaluate the accuracy and speed of various matching software packages on test decks provided by the Decennial IT Division.

Staff: William Winkler (x34729), Emanuel Ben-David, Ned Porter

B. Coverage Measurement Research

Description: Staff members conduct research on model-based small area estimation of census coverage, and they consult and collaborate on modeling census coverage measurement (CCM).

Highlights: During FY 2016, staff attended meetings to discuss and develop the sampling and estimation procedures needed for the 2020 Census Coverage Measurement program. The staff helped make recommendations on which topics warrant further research for possible implementation for 2020.

Staff: Jerry Maples (x32873), Ryan Janicki, Eric Slud

C. Using 2010 Census Coverage Measurement Data to Compare Nonresponse Follow-up Proxy Responses with Administrative Records

Description: Research in preparation for the 2020 Census Nonresponse Follow-up (NRFU) investigates employing different contact strategies along with the use of administrative records (AR) files to reduce the cost of the operation while maintaining data quality. Regardless of the contact strategy, one asks whether the proxy responses are more accurate than ARs available for the NRFU housing units (HUs). The goal of this study is to use the results of the 2010 Census Coverage Measurement Program (CCM) to compare the accuracy of proxy responses for 2010 Census NRFU housing units in the CCM sample with the accuracy of the ARs available for the housing units.

Highlights: During FY 2016, the investigation discovered that the percentage of enumerations in HUs with proxy respondents in the correct location in the CCM HUs was higher than the percentage of ARs in the same HUs even though the AR sources were all IRS 1040 and Medicare records from 2010. However, the percentage of records that could not be evaluated was higher for the ARs than for the proxy respondents. The high unresolved rate among ARs was due to the failure to link the AR to a combined CCM record at the same address. The reasons that an AR did not link include the individual being enumerated at another address, having a census enumeration or P-sample roster entry that could not be assigned a Protected Identification Key (PIK), or being missed by the census. This research prompted a change from the initial plan that used all ARs for NRFU enumeration to the search for methods to identify the best ARs for enumeration. The current methodological approach focuses on the development of predictive models to identify ARs with a high probability of being accurate. Staff submitted a report containing the results to the *CSRM Research Report Series*.

Staff: Mary Mulry (x31759)

D. Record Linkage Error-Rate Estimation Methods

Description: This project develops methods for estimating false-match and false-nonmatch rates without training data and with exceptionally small amounts of judiciously chosen training data. It also develops methods/software for adjusting statistical analyses of merged files when there is linkage error.

Highlights: Staff worked on automatic error-rate estimation for record linkage for more than fifteen months. The staff included individuals from CSRM, the Decennial Statistical Studies Division (DSSD), and the Center for Administrative Records Research and Applications (CARRA). More recently, this project has been on hold. This project may begin in FY 2017 as follow-on of evaluation and testing of record linkage packages for DSSD.

Staff: William E. Winkler (x34729), Emanuel Ben-David, Tom Mule (DSSD)

E. Supplementing and Supporting Non-Response with Administrative Records

Description: This project researches how to use administrative records in the planning, preparation, and implementation of nonresponse follow-up to significantly reduce decennial census cost while maintaining quality. The project is coordinated by one of the 2020 Census Integrated Project Teams.

Highlights: During FY 2016, staff continued to analyze the results of stepwise logistic regression models on Nonresponse Follow-up (NRFU) IDs in Maricopa County, Arizona, with topcoded Census Unedited File (CUF) household size as the dependent variable. Staff fit models on a random 5% subsample of the Maricopa file and scored the models on the entire file. The fitting and scoring were both done in four separate pieces: Undeliverable As Addressed (UAA) flag blank and tax year 2009 IRS 1040 household count at least one, UAA flag blank and tax year 2009 IRS 1040 household count zero but tax year 2008 IRS 1040 household count at least one, UAA flag blank and both tax year 2008 and tax year 2009 IRS 1040 household counts zero, UAA flag nonblank. The results from the four pieces were combined into a single output file. The scored output (including predicted probabilities, predicted household size, and input model variables) was provided to staff in the Decennial Statistical Studies Division (DSSD). DSSD used the predicted probabilities to incorporate constraints on the expected value of household size into models for using administrative records (AR) for occupied housing units (HUs). Adding constraints on expected household size showed some promise for helping to maintain the overall population count.

Staff also fit models on a random 5% subsample of a national NRFU file and scored the models on the entire file. As with the Maricopa file, the fitting and scoring were both done in four separate pieces: UAA flag blank and tax year 2009 IRS 1040 household count at least one, UAA flag blank and tax year 2009 IRS 1040 household count zero but tax year 2008 IRS 1040 household count at least one, UAA flag blank and both tax year 2008 and tax year 2009 IRS 1040 household counts zero, UAA flag nonblank. The results from the four pieces were combined into a single output file. The scored output (including predicted probabilities, predicted household size, and input model variables) was provided to DSSD. As with the Maricopa file, DSSD used the predicted probabilities to incorporate constraints on the expected value of household size into models for using AR for occupied HUs. Adding constraints on expected household size again showed some promise for helping to maintain the overall population count. Staff also compared several different models on the group of IDs where the UAA flag is blank and the tax year 2009 IRS 1040 household count is at least one. For these IDs, the distribution of the expected value of household size (by topcoded tax year 2009 IRS 1040 count) and the distribution of the rounded expected value of household size are reasonably similar across models. However, separate models for each value of topcoded IRS household count are needed for the distribution of the predicted value (household size with the maximum estimated probability) to be similar to the distribution of topcoded CUF household size. Staff also looked at the effect of using cutoffs for the maximum estimated probability in determining when to use the household count based on administrative records. This part of the analysis included IDs where the UAA flag was blank and either the tax year 2009 IRS 1040 household count or the tax year 2008 IRS 1040 household count was at least one. Separate models were fit for each value of the 2009 IRS 1040 household count when that count was at least one, otherwise separate models were fit for each value of the 2008 IRS 1040 household count. IDs with predicted values other than 1-6 were excluded, and quartiles of the maximum estimated probability were calculated based on the remaining IDs and used as cutoffs. As the cutoff got more restrictive, the rounded expected value and the predicted value became more concentrated on household sizes of one, two, and four (especially one) and also showed fewer differences from each other. For the predicted value, using no probability cutoff resulted in an overestimate compared to the CUF, while using the third quartile as the cutoff resulted in an underestimate. For the rounded expected value, the estimates are similar to the CUF results both for no cutoff and for cutoffs at any quartile. The results are similar when IDs enumerated on the first contact are ineligible for using the modeled count, although the distributions of the remaining rounded expected values, predicted values, and CUF household counts shift a bit towards smaller counts. The ratio of the AR-based estimates to the CUF results increases somewhat (compared to the corresponding ratios including IDs enumerated on the first contact) for those units that meet the cutoffs. However, since there are fewer of these units the net effects on the NRFU estimates are similar. Staff summarized the results of the analyses of the modeling on the national data file in a draft document that was sent to a subgroup of the Administrative Records Modeling Team.

Staff also attended team meetings and reviewed work by other team members.

Staff: Michael Ikeda (x31756), Mary Mulry

F. Identifying "Good" Administrative Records for 2020 Census NRFU Curtailment Targeting

Description: As part of the Census 2020 Administrative Records Modeling Team, staff are researching scenarios of nonresponse follow-up (NRFU) contact strategies and utilization of administrative records data. Staff want to identify scenarios that have reduction in NRFU workloads while still maintaining good census coverage. Staff are researching identification of "good" administrative records via models of the match between Census and administrative records person/address assignments for use in deciding which NRFU households to continue to contact and which to primary allocate. Staff are exploring various models, methods, and classification rules to determine a targeting strategy that obtains good Census coverage—and good characteristic enumeration—with the use of administrative records.

Highlights: During FY 2016, staff published "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Decennial Census" in the June issue of the *Statistical Journal of the International Association of Official Statistics* as part of a series of papers on using administrative records in the 2020 Census. This paper documents a scenario of administrative records vacancy and occupancy (enumeration) determination using a linear programming approach. Staff finalized the paper "A Modeling Approach for Administrative Record Enumeration in the Decennial Census" accepted for a special issue of *Public Opinion Quarterly*. This paper compares classification methods for a person-place model for administrative records usage. Staff investigated decision theoretical strategies to exploit information on response propensity, along with information on administrative records (AR) quality, when conducting NRFU. Staff presented and documented findings at the 2016 Joint Statistical Meetings in Chicago and in a *JSM Proceedings* paper titled "Bayesian Decision Theory to Optimize the Use of Administrative Records in Census NRFU." These findings include theoretical results in decision theory as well as general operational scenarios for Census 2020. Staff continued to attend meetings and provide input into research topics studied by the administrative records modeling team such as comparison of models using 2010 Census vs. ACS data, an alternative approach for determining administrative records removals using a distance function based on predicted probabilities, and the analysis of USPS information in the 2016 Census test.

Staff: Darcy Steeg Morris (x33989), Yves Thibaudeau

G. Evaluation of Response Error Using Administrative Records

Description: Censuses and their evaluations ask respondents to recall where they lived on Census Day, April 1. Some interviews for evaluations take place up to eleven months after this date. Respondents are asked when they moved to their current address, and the assumption has been that respondents who move around April 1 are able to give correct answers. Error in recalling a move or a move date may cause respondents to be enumerated at the wrong location in the census. This study investigates recall error in reports of moves and move dates in censuses and sample surveys using data from survey files linked to administrative records.

Highlights: During FY 2016, staff continued to collaborate with staff in the Center for Survey Measurement (CSM) on analyses of recall error for reports of moves and move dates in surveys using data from survey files linked to administrative records. Staff pursued two studies. One study uses data from the Recall Bias Study, which was part of the 2010 Census Evaluation and Experiments Program. Results from the study were published by the refereed journal *Survey Methods: Insights from the Field* in July. The other study uses data prepared for the "Memory Recall of Migration Dates in the National Longitudinal Survey of Youth" developed under a contract with the National Opinion Research Center (NORC). Staff continues to improve the draft manuscript for this study by addressing the comments received from reviewers. In addition, staff presented an invited paper regarding lessons learned about evaluating survey data with administrative records files at the 2016 Methodology Symposium sponsored by Statistics Canada and submitted an invited paper to the *Proceedings of the 2016 Methodology Symposium*.

Staff: Mary Mulry (x31759)

H. Special Census: Disclosure Avoidance in Group Quarters

Description: Staff works with the Decennial Information Technology Division (DITD) to create synthetic data for disclosure avoidance in group quarters data for ongoing Special Census production.

Highlights: During FY 2016, staff worked with DITD in creating synthetic data for disclosure avoidance in group quarters data for ongoing Special Census production for certain localities in Iowa, Illinois, and Arizona. Staff determined which data were at potential risk of disclosure and applied statistical models to produce new data to replace those items. DITD integrated this data into the final product. Work on this project is now complete.

Staff: Rolando Rodriguez (x31816)

I. 2020 Unduplication Research

Description: The goal of this project is to conduct research to guide the development and assessment of methods for conducting nationwide matching and unduplication in the 2020 Decennial Census, future Censuses, and other matching projects. Our staff will also develop and test new methodologies for unduplication. The project is coordinated by one of the 2020 Census Integrated Project Teams.

Highlights: During FY 2016, staff continued to investigate networked links to identify unusual coincidental matches.

Staff: Michael Ikeda (x31756), Ned Porter, Bill Winkler, Emanuel Ben-David

J. Analysis of the 2015 Census Test Evaluation Follow-up

Description: The U.S. Census Bureau conducted the 2015 Census Test as part of its research to develop methodology for using administrative records (ARs) to reduce the cost and improve the quality of the 2020 Census Nonresponse Follow-up (NRFU) data. The goal of the 2015 Census Test in Maricopa County, AZ was to test methodology and operations designed to reduce the NRFU workload. The 2015 Evaluation Follow-up (EFU) was part of the 2015 Census Test and collected additional data to allow a comparison of NRFU data with ARs available for the same addresses. The 2015 EFU analyses provide information about different uses of ARs that are topics of current research, such as determining occupancy status, enumerating a housing unit (HU), and providing data for imputation procedures. The 2015 EFU interviewed 4,098 HUs where there was a discrepancy between the NRFU results and the ARs.

Highlights: During FY 2016, staff collaborated with Decennial Statistical Studies (DSSD) staff on the analysis of data collected in the 2015 EFU. The team performed separate analyses for NRFU household (HH) member respondents and proxy respondents using addresses where both NRFU and ARs agreed that they were occupied but the population counts differed. For HH member respondents, the EFU interview results agreed with the NRFU results 55.8% of the time and agreed with the administrative record count 17.5% of the time. For 21.7%, EFU provided a different count than both ARs and NRFU. The remaining 5.0% had an unresolved status in EFU. In addition, the EFU HH composition agreed with the EFU HH composition at a higher rate than with the AR HH composition. This result led to the team's recommendation of including an additional mailing to addresses without a self-response but with ARs that appeared of high enough quality for enumeration. The recommended mailing was implemented in the 2016 Census Test.

For NRFU proxy respondents, EFU agreed with the administrative record count 32.5% of the time, agreed with the NRFU count 33.4% of the time and disagreed with both 22.3% of the time. Further analysis indicated that the characteristics were missing fewer times for the ARs than the NRFU proxy respondents. The EFU results indicated that ARs appear to be of comparable quality to proxy enumerations for the count; for characteristics, ARs may be better. Additional support came from the observation that the characteristics were missing fewer times for the administrative records where processing was able to assign a Protected Identification Key than for the NRFU proxy respondents. These results supported the continuation of planned research on using ARs instead of proxy responses for enumerating addresses where high quality administrative records are available.

Staff also completed an internal memorandum that contains the results of the analysis. Staff presented the results at the 2016 Joint Statistical Meetings and submitted a paper to the *JSM Proceedings*.

Staff: Mary Mulry

1.5 ADDRESS CANVASSING IN FIELD (Decennial Project 6350B02)

A. Master Address File (MAF) Error Model and Quality Assessment
Description: The MAF is an inventory of addresses for all known living quarters in the U.S. and Puerto Rico. This project will develop a statistical model for MAF errors for housing units (HUs), group quarters (GQs), and transitory locations (TLs). This model, as well as an independent team, will be used to conduct independent quality checks on updates to the MAF and to ensure that these quality levels meet the 2020 Census requirements.

Highlights: During FY 2016, staff completed a research report on zero-inflated negative binomial modeling with exhaustive variable selection using 2010 address canvassing database plus several supplemental data sources. Under the selected model, only a small number of blocks exhibited large residuals; however, many of these blocks coincide with large add counts. Staff applied spatial generalized mixed models to selected regions (Baltimore and New York counties) to determine the utility of accounting for spatial dependence. Following a Bayesian approach of Hughes and Haran (2013) to reduce the dimension of the spatial effects, only subtle improvements are seen in predictions over nonspatial models while regression coefficients and their significance change substantially. Spatial correlation in residuals is also reduced in spatial models. Staff considered the use of statistical decision theory to aid decisions such as "canvass" versus "do not canvass" when the underlying state of coverage error in a block takes on categories such as "Low", "Medium", and "High". Initial work emphasizes that optimal decisions are sensitive to the decision maker's choice of utility function.

Staff: Andrew Raim (x37894), Laura Ferreira (DSSD), Krista Heim (DSSD), Scott Holan (University of Missouri)

B. Development of Block Tracking Database
Description: The Targeted Address Canvassing (TRMAC) project supports Reengineered Address Canvassing for the 2020 Census. The primary goal of the TRMAC project is to identify geographic areas to be managed in the office (i.e., in-office canvassing) and geographic areas to be canvassed in the field. The focus of the effort is on decreasing in-field and assuring the Master Address File (MAF) is current, complete, and accurate. The Block Assessment, Research, and Classification Application (BARCA) is an interactive review tool which will allow analysts to assess tabulation blocks—and later Basic Collection Units (BCUs)—by comparing housing units in 2010 imagery and current imagery, along with TIGER reference layers and MAF data.

Highlights: During FY 2016, staff designed a quality control feature into BARCA that enables an adjudicator to assess any reviewer errors and correct them before results are loaded into the BTD. A system of reports has also been added to the system. The In-Office Address Canvassing Interactive Review also reached the FY16 Goal of completing 50% of all blocks in the US and Puerto Rico.

Staff: Tom Petkunas (x33216)

C. Detection of Map Changes
Description: This research is concerned with developing statistical techniques to detect changes in maps, utilizing remote sensing data, such as LIDAR.

Highlights: During FY 2016, staff (a) collaborated with Census geographers to develop methods for detecting map changes; (b) learned GIS (Geographic Information System) software to process data; and (c) improved road detection algorithms by adding a connectivity criterion to identify road networks. Staff also wrote Matlab code for a modified Hough transform line detector, with results comparable to shearlet methodology.

Staff completed an investigation of the history and goals of the in-office canvassing problem. From discussions with Geography Division staff, it was determined that Master Address File (MAF) errors can occur by many distinct processes, e.g., conversion of a single family home to a multi-unit, addresses from the Delivery Sequence File (DSF) left ungeocoded, and past human error. Restricting the scope of the problem will be necessary in order to proceed further.

Staff: Dan Weinberg (x38854)

1.6 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 6385B70)

A. ACS Applications for Time Series Methods
Description: This project undertakes research and studies on applying time series methodology in support of the American Community Survey (ACS).

Highlights: During FY 2016, staff extended R code for custom multi-year estimates to handle large data frames and addressed input-output issues. Staff met with clients from the Veterans Administration (VA) to discuss practical aspects of the project such as how the format of the data is prepared for us and the desired output.

Staff also extended methodology and R code for custom multi-year estimates to handle point estimates, as opposed to period estimates. Staff refined code and preliminary results for all counties. Staff explained the interpolation methodology through presentations to Census and VA staff. A final draft of the technical paper was completed.

Staff: Tucker McElroy (x33227), Osbert Pang

B. Data Analysis of ACS CATI-CAPI Contact History
Description: This project continues earlier analyses of the American Community Survey (ACS) Computer Assisted Telephone Interview (CATI) and Computer Assisted Personal Interview (CAPI) contact history data. It focuses exclusively on CAPI with the goal of informing policy decisions on curtailing of CAPI contact attempts to minimize respondent burden on sampled households without unacceptable losses of ACS interviews.

Highlights: During FY 2016, staff completed an analysis of the results of an ACS Pilot Study conducted in August 2015 on one month’s ACS data collections in 12 of 48 Supervisory Field Areas nationally. The Pilot had been designed to assess the effects of a new policy curtailing field nonresponse follow-up on ACS personal-interviewer (CAPI) cases when a measure of “cumulative burden” experienced by a household exceeded a pre-defined threshold, in anticipation of national rollout of this policy. The goal of the policy was to reduce burden on potential respondents without major decreases in the CAPI case rate of interview completion. The policy had been developed based on earlier research in this project, and consisted of withdrawing a case from its field representative's (FR's) workload when a measure of the cumulative burden imposed on a potential respondent as a result of repeated contact attempts (including attempts in pre-CAI modes) crossed a pre-set “maximum burden” threshold.

Staff analyzed the pilot study results with the objective of comparing lost interviews and workload reduction with the levels forecast from analysis of previous (2012) ACS data, and of testing the comparative outcomes of workload and lowered interview completion rate between three “treatments”. The treatments consisted of implementation of the policy with and without telling the FRs each day what their current cumulative-burden score was and a control in which cases were not removed as a result of exceeding the cumulative-burden threshold. The Pilot was analyzed as far as possible as an embedded designed experiment, for comparison with projected results (from earlier research on 2012 ACS data) on workload, interview completion rates, and burden experienced.

The research resulted in an *ACS Research and Evaluation Series Report*. Staff wrote up this research more broadly in a survey of the ACS’ research program to reduce respondent burden, in a submitted journal paper.

Staff: Eric Slud (x34991), Robert Ashmead, Todd Hughes (ACSO), Rachael Walsh (OSCA)

C. Assessing Uncertainty in ACS Ranking Tables
Description: This project presents results from applying statistical methods which provide statements of how good the rankings are in the ACS Ranking Tables (see The Ranking Project: Methodology Development and Evaluation Research Section under Projects 0331000 and 0925000).

Highlights: [See General Research: Survey Sampling-Estimation and Modeling (C), The Ranking Project: Methodology Development and Evaluation]

Staff: Tommy Wright (x31702), Martin Klein, Jerzy Wieczorek (Carnegie Mellon University), Brett Moran, Nathan Yau, Michael Leibert

D. Confidence Intervals for Proportions in ACS Data
[See General Research: Small Area Estimation (B), Coverage Properties of Confidence Intervals for Proportions in Complex Surveys]

E. Voting Rights Section 203 Model Evaluation and Enhancements Towards Mid-Decadal Determinations
Description: Section 203 of the *Voting Rights Act* mandates determinations by the Census Bureau relating to rates of citizenship, limited English proficiency and limited education among specified small subpopulations (voting-age persons in various race and ethnicity groups called Language Minority Groups [LMGs] for small areas such as counties or minor civil divisions MCDs). The Section 203 determinations for 2016 will result in the legally enforceable requirement that certain geographic political subdivisions must provide voting materials in languages other than English in future elections. The research undertaken in this project consists of the development, assessment and estimation of regression-based small area models based on 5-year (2010-2014) American Community Survey (ACS) data. This modeling and estimation effort differs from the effort supporting Section 203 determinations in 2011. The 2016 models and estimates cannot make use of data from a nearly contemporary decennial census.

Highlights: Under the general guidance of the Census Redistricting & Voting Rights Data Office and the Decennial Statistical Studies Division (DSSD), staff worked throughout FY 2016 to develop modeling techniques and estimates for the voting populations and subpopulations of citizens, limited English proficiency (LEP) citizens and illiterate LEP citizens within LMGs and political subdivisions including states, counties, MCDs and American Indian Areas (AIAs). These small subpopulations are here referred to as small domains of the US population.

Staff first performed descriptive and exploratory data analyses using variants of the empirical-Bayes (specifically, beta-binomial) models used in 2011 to provide analogous estimates, with the new models based only on five-year ACS 2008-2012 data sources. Early findings included the high variability of direct survey-weighted estimates from ACS data, implying the necessity of using models to provide small-domain estimates at least in the smaller jurisdictions (counties and MCDs) and AIAs as well as the benefits of using synthetic (state-level) and small-domain covariates as predictor variables in empirical-Bayes regression models. The models found to be most promising were Dirichlet-multinomial regression models with covariates including state-level rates of citizenship and LEP within the voting-age population as well as small-domain-level covariates summarizing educational level, average length of time spent in US, proportion foreign-born, and age, both for the small-domain adult population as a whole and also for the single LMGs being modeled.

Staff conducted further investigations concerning the choice of predictor variables to include in models as a function of the numbers of jurisdictions or AIAs with ACS five-year sample within each LMG and the convergence of maximum likelihood estimates within reasonable parameter ranges based on the resulting regression models. Staff programmed the small-domain estimates based on Dirichlet-multinomial regression models for all 68 LMGs and political units. Staff also devised a method for the estimation of variances of the resulting estimates based on a novel hybrid approach combining replicate weights as used in ACS with parametric-bootstrap generation of pseudo-samples of survey data.

Staff prepared preliminary documentation in the spring and early summer of 2016 supporting the modeling choices made and briefed officials in the Director's Office on the basis of those documents. Staff ran the models on 2010-2014 data and delivered point and variance small-domain population estimates to DSSD and the Census Redistricting & Voting Rights Data Office in Fall 2016 after first conducting extensive quality-control and programming checks on the resulting data products. Work is continuing on complete technical documentation of the estimation methodology and the assessments made of the preceding modeling choices.

Staff: Patrick Joyce (x36793), Robert Ashmead, Eric Slud, Mark Asiala (DSSD)

1.7 DEMOGRAPHIC STATISTICAL METHODS DIVISION SPECIAL PROJECTS (Demographic Project TBA)

A. Special Project on Weighting and Estimation
Description: This project involves regular consulting with Current Population Survey (CPS) Branch staff on design, weighting, and estimation issues regarding the CPS. Issues discussed include design strategy for systematic sampling intervals, for rotating panels, composite estimation, variance estimation, and the possibility of altering CPS weighting procedures to allow for a single simultaneous stage of weight-adjustment for nonresponse and population controls.

Highlights: No significant updates during FY 2016.

Staff: Eric Slud (x34991), Yang Cheng (DSMD)

B. Weighted Estimating Equations with Response Propensities in Terms of Covariates Observed Only for Responders
Description: The project now considers not only survey response propensities but also data on an administrative-records observational database, with the goal of modeling joint indicators of survey response and administrative-list inclusion. Staff aims to develop survey analysis methods incorporating administrative data that can, under suitable model assumptions, provide representative population estimators.

Highlights: During FY 2016, staff presented a talk on this topic at a National Institute of Statistical Sciences Workshop on Nonignorable Nonresponse on November 12-13, 2015 and began a series of meetings and collaborative discussion on applying Weighted Estimating Equations to the problem of design and analysis of simultaneous survey and administrative-list data. This research culminated in a literature review and simulation study illustrating the benefit of jointly analyzing such data when available, and a talk, “Design of Sample Surveys That Complement Observational Data to Achieve Population Coverage”. It was delivered as a Contributed paper at the 2016 Joint Statistical Meetings.

Staff: Eric Slud (x34991), Robert Ashmead, Anindya Roy

1.8 DEMOGRAPHIC SURVEYS DIVISION (DSD) SPECIAL PROJECTS (Demographic Project 0906/1444X00)

A. Data Integration

Description: The purpose of this research is to identify microdata records at risk of disclosure due to publicly available databases. Microdata from all Census Bureau sample surveys and censuses will be examined. Potentially linkable data files will be identified. Disclosure avoidance procedures will be developed and applied to protect any records at risk of disclosure.

Highlights: During FY 2016, staff worked with the Center for Disclosure Avoidance Research (CDAR) to continue planning stages to confirm suspected records for re-identification in the American Housing Survey (AHS). These stages include outlining software specifications for confirming suspicious records, identifying input file variables, and developing weighted likelihood measures of re-identification for the output file. Test runs of a beta algorithm on simulated data showed efficiency of the algorithm. CDAR published the results of the AHS re-identification study internally. Staff also developed software to clean data for the American Community Survey and the 2000 Decennial Census Long Form to support differential privacy research. The project has also been expanded to include Decennial Census Data.

Staff: Ned Porter (x31798), Marlow Lemons (CDAR)

1.9 POPULATION DIVISION PROJECTS (Demographic Project TBA)

A. Introductory Sampling Workshop

Description: In support of Population Division's International Programs Area, staff will conduct (on request) introductory sampling workshops with focus on probability sampling for participants from various countries. These workshops are primarily funded by USAID.

Highlights: Over the two-week period (October 26-November 6, 2015), staff conducted a Workshop: Introduction to Survey Sampling (focus on Probability Sampling) at the Census Bureau Headquarters. The workshop presented the main components of survey sampling with a focus on probability sampling (and estimation) techniques. The hands-on, interactive workshop included the production of estimates of population parameters from sample surveys as a function of sample design, weighting procedures, the computation of sampling errors of sample estimators, and the making of inferences from the sample to the population. The seven workshop participants were mostly staff from statistical agencies in the United States, Ethiopia, Angola, and Namibia.

On the final day, the workshop featured a Panel on Sampling to give overviews of the American Community Survey, the Monthly/Annual Retail Trade Surveys, and the Current Population Survey.

Plans are in place to offer the workshop in the Fall 2016 for international participants.

Staff: Tommy Wright (x31702), Michael Leibert

1.10 SOCIAL, ECONOMIC, AND HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165016)

A. Research for Small Area Income and Poverty Estimates (SAIPE)

Description: The purpose of this research is to develop, in collaboration with the Small Area Estimates Branch in the Social, Economic, and Housing Statistics Division (SEHSD), methods to produce "reliable" income and poverty estimates for small geographic areas and/or small demographic domains (e.g., poor children age 5-17 for counties). The methods should also produce realistic measures of the accuracy of the estimates (standard errors). The investigation will include assessment of the value of various auxiliary data (from administrative records or surveys) in producing the desired estimates. Also included would be an evaluation of the techniques developed, along with documentation of the methodology.

Highlights: During FY 2016, staff explored methodology for the estimation of year-to-year changes in poverty rates for children through jointly modeling two years of ACS one-year estimates through a bivariate Fay-Herriot model with measurement error, where the measurement error variable comprises past estimates of children in poverty derived from ACS five-year data. The results were compared with results from a bivariate model that excluded the five-year estimates, and with a naïve model that treats the ACS five-year estimates as a covariate with no measurement error. These models were applied to the ACS estimates both with and without additional covariates from administrative records to understand the impact of these covariates. Staff implemented all of these models in a Bayesian setting initially using the prior distributions in the generic software JAGS. Staff also developed a Markov Chain Monte Carlo (MCMC) algorithm tailored to this model, using a class of priors for which staff proved the propriety of the posterior under mild conditions. Different implementations of the model were explored as well as different prior distributions for the parameters and different parametrizations. Algorithms developed for this model included an MCMC algorithm that uses a combination of Gibbs Sampling and Metropolis Hastings, as well as an importance-sampling algorithm used for comparison. Staff wrote a paper on this work, which is currently under revision for a peer reviewed journal.

Staff also performed several model diagnostics to determine whether the bivariate BLN model discussed in Franco and Bell (2015) could perform better than the current production model for rates of school-aged children in poverty at the county-level. Staff compared aggregations of the BLN model and production model under partitions of the sample size such that each element of the partition is large enough to have negligible sampling error in the direct estimates. Staff found that when partitioning by population size, sample size, and by each of the four model covariates, the aggregates of the BLN model match those of the direct more closely than those of the current production model without raking. The BLN model's aggregate estimate for the nation is also closer to that of the current production before raking. These results hold consistently for the three years examined (2010-2012). The results also show a potential bias for the production model for small sample sizes.

Staff: Jerry Maples (x32873), Carolina Franco, William Bell (ADRM)

B. Small Area Health Insurance Estimates (SAHIE)
Description: At the request of staff from the Social, Economic, and Housing Statistics Division (SEHSD), our staff will review current methodology for making small area estimates for health insurance coverage by state and poverty level. Staff will work on selected topics of SAHIE estimation methodology, in conjunction with SEHSD.

Comparing Small Area Estimates over Time
Highlights: During FY 2016, staff developed models for inference on year-to-year change of small area parameters from different time periods using beta mixed effects regression models. For marginal beta sampling distributions with Gaussian model errors and a logistic link function, staff proved propriety of the posterior distribution using a noninformative prior, proved convergence of the posterior mean to either the direct estimate or a synthetic estimate as either the model variance or the direct variance tends to zero (a shrinkage property), and developed a Markov Chain Monte Carlo (MCMC) algorithm for sampling from the posterior distribution. This model was extended to incorporate data from multiple time periods using multivariate Gaussian model errors. For the multivariate model, different transformations and series expansions were investigated to obtain moment estimators for the fixed effects and variance components. Simulation studies were conducted to understand properties of these estimators.

Staff: Ryan Janicki (x35725)

C. Sub-County Estimates of Poverty from Multi-year ACS Data
Description: This project is from the Development Case Proposal to improve the estimates of poverty related outcomes from the American Community Survey (ACS) at the tract level. Various modeling techniques, including model-based and model-assisted, will be used to improve on the design-based multi-year estimates currently produced by the ACS. The goal is to produce more accurate estimates of poverty and income at the tract level and develop a model framework that can be extended to outcomes beyond poverty and income.

Highlights: During FY 2016, staff investigated models for the number of poor school-aged children in tracts using three different variants of over-dispersed Poisson models: CMP (Conway and Maxwell), Generalized Poisson (Consul and Jian) and Double Exponential family for Poisson (Efron), in addition to the Negative Binomial distribution. The different distributions will be tested against the artificial population data samples. These over-dispersed count distributions will then be used to form an area-level small area model to predict the number of school-age children in poverty. The purpose of this project is to compare estimates against rate models that assume a known population total at the sub-county level.

Staff: Jerry Maples (x32873), Ryan Janicki, Carolina Franco, William Bell (ADRM)

1.11 ECONOMIC STATISTICAL COLLECTION (Economic Project 1183X01)

1.12 ECONOMIC MONTHLY/RETAIL (Economic Project 1001X00)

A. Research on Imputation Methodology for the Monthly Wholesale Trade Survey
Description: In the previous phase of this project, staff conducted a simulation study to investigate new imputation methodology for the Monthly Wholesale Trade Survey (MWTS). In this phase of the project, staff are creating a more realistic simulated wholesale trade population and investigating improved MWTS estimators. The MWTS is a longitudinal survey that provides month-to-month information on sales and inventories of U.S. merchant wholesalers. Key estimates produced from this survey include total sales, month-to-month relative change in sales, total

inventories, and month-to-month relative change in inventories (overall and within industry subclasses). There are a number of challenges when developing estimators for the MWTS, including variables with highly skewed distributions, missing values in predictor variables from the Economic Census, and survey variables with trends that differ across industry classes. The longitudinal information in addition to a rich set of frame data available from the Economic Census can be used to build Bayesian models that address these challenges. It is expected that this model will be applicable to other business surveys.

Highlights: During FY 2016, staff developed a realistic, artificial population that can be used to repeatedly draw simulated Monthly Wholesale Trade Survey (MWTS) data representative of the two-year period from December 2008 to December 2010. The constructed population enables us to study the statistical properties of new estimation and imputation procedures on the MWTS. Staff first evaluated a version of this population that was previously developed in order to identify areas where improvements could be made and new features could be incorporated. Staff then used sampling frame information, auxiliary data, and MWTS data from December 2008 to December 2010 to completely redesign and construct a new version of the population. Staff developed the population in two parts: (1) the part consisting of units that are sampled with certainty, and (2) the part consisting of the remaining units which are sampled based on stratified random sampling design. To construct the certainty population, staff used observed MWTS data over a two-year period to obtain a roster of sampling units, reporting units, and tabulation units. These units were then merged with data from the sampling frame and other auxiliary data. Missing values in the certainty population were then imputed using multivariate imputation by chained equations with random forests as the conditional models. Staff experimented with variations of this imputation method and analyzed the outputs in order to find a good imputation model. To construct the population of non-certainty units, staff used data from the sampling frame to obtain a roster of units and then merged it with MWTS and auxiliary data. The next step was to fill in any missing data to get a complete non-certainty population. As with the certainty population, staff used multivariate imputation by chained equations with random forests as the conditional models to impute missing data in the non-certainty population. As with the certainty population, staff used multivariate imputation by chained equations with random forests as the conditional models to impute missing data in the non-certainty population. However, since many non-certainty units are not actually in the MWTS sample, a very large proportion of the non-certainty population have all values missing for the MWTS variables. This large scale imputation poses some challenges; for instance, if we used a standard random forest procedure to impute all the non-sampled units, then the imputed population would contain a relatively small number of repeating values. Thus, to avoid having the same values repeated many times in the population, staff customized a random forest imputation procedure so that it draws the imputed value from a kernel density fitted to the candidate values, instead of drawing the imputed value directly from the candidate values. Staff used this customized imputation procedure to impute missing data for non-certainty units that were not in the MWTS sample. Analysis of the newly constructed population indicates that it presents a realistic portrayal of the true population, especially with respect to month-to-month percent change in sales and inventories, which is an important feature. Another desirable feature of the population is that it incorporates information about companies that operate in multiple industry classes. This population can be used to repeatedly draw simulated MWTS data, allowing one to design simulation studies to evaluate new statistical methodology for the survey.

Staff: Martin Klein (x37856), Joe Schafer (ADRM), Joanna Fane Lineback (CSM), Brett Moran

B. Use of Big Data for Retail Sales

Description: In this project, we are investigating the use of “Big Data” to fill gaps in retail sales estimates currently produced by the Census Bureau. Specifically, we are interested in how to use “Big Data” to supplement existing monthly/annual retail surveys with a primary focus on exploring (1) how to use third party data to produce geographic level estimates more frequently than once every five years (i.e. a new product), and (2) the possibility of using third party data tabulations to improve/enhance Census Bureau estimates of monthly retail sales - for example, validation and calibration. Various types of data are being pursued such as credit card transaction data and scanner data.

Highlights: During FY 2016, staff contributed to a final report and internal presentations of the evaluation of the quality of aggregated data from a major credit card. Staff worked with Economic Directorate staff to assess the quality and potential use of aggregated electronic transaction data from First Data. Staff participated in working meetings with staff from Palantir, the company that houses, manages and visualizes the First Data data. Staff studied and contributed to a report on small area estimation models for assessing the predictive power of the First Data transaction data for industry/state level estimates of monthly change and retail sales totals. Staff continued to provide general input and suggestions for the visualization and computing tool that Palantir developed for analysis of data from First Data.

Staff: Darcy Steeg Morris (x33989), Osbert Pang, Tommy Wright, Bill Bostic (ADEP), Scott Scheleur (SSSD), Rebecca Hutchinson, Bill Davie, Jr. (ESMD)

1.13 ECONOMIC CENSUS/SURVEY ENGINEERING: TIME SERIES RESEARCH; ECONOMIC MISSING DATA/PRODUCT LINE DATA; DEVELOPMENT/SAS (Economic Project 2220B10)

A. Seasonal Adjustment Support

Description: This is an amalgamation of projects whose composition varies from year to year but always includes maintenance of the seasonal adjustment software used by the Economic Directorate.

Highlights: During FY 2016, staff provided seasonal adjustment and software support for users within and outside the Census Bureau, including EJJE (Mexico), Epistemic, Ernst and Young, Goldman Sachs, M&T Bank, Nikkei Inc. (Japan), Obiettivo Lavoro (Italy), Palantir, Vanguard, SAS, Conference Board, Credit-Suisse, Reserve Bank of New Zealand, New York Federal Reserve Board, Bureau of Economic Analysis, Bureau of Labor Statistics, European Central Bank, Swiss National Bank, KOF Swiss Economic Institute (Switzerland), National Bureau of Statistics (Nigeria), INSEE (France), Office of National Statistics (UK), Statistics Canada, Statistics New Zealand, Statistics Centre of Abu Dhabi, INDEC (Argentina), Quebec Institute of Statistics, Department for Environment, Food and Rural Affairs (UK), Statistics Norway, Barcelona Graduate School for Economics, University of Bern, Columbia University, Northwestern, Catholic University of Louvain, and Zhejiang University.

Staff participated in a briefing on seasonal adjustment practice at the Census Bureau given to Mr. Silvio Peruzzo of Rokos Capital Management LLP on behalf of the Italian government.

Staff organized and participated in a meeting with Economic Directorate staff to discuss residual seasonality in the Value Put in Place series. Staff participated in a time series workshop in London, UK on November 19 and 20, 2015, sponsored by the Office of National Statistics. Staff presented the current state of time series research and development at the Census Bureau and participated in brainstorming sessions on future time series work.

Staff met with Ronald Indergand of the University of Bern to discuss his work comparing seasonal adjustment revisions from X-11 and SEATS. Staff worked with other Census staff to answer questions for a *Wall Street Journal* reporter doing an article on seasonal adjustment.

Staff organized a 2016 *Summer at Census* visit by Christophe Sax, a consultant who has developed the seasonal R package that interfaces with X-13ARIMA-SEATS. Sax gave a talk and had discussions with Census Bureau personnel on future work related to the seasonal package. Staff participated in a briefing on seasonal adjustment practice at the Census Bureau to Mr. Silvio Peruzzo of Rokos Capital Management LLP on behalf of the Italian government.

Staff met with Center for Economic Studies (CES) staff on a plan to produce seasonal adjustments of various quarterly series (both state and national level) that are not currently being seasonally adjusted.

Staff began meeting with an interagency seasonal adjustment group to study and remediate residual seasonality in GDP and other series. Staff met to discuss participation in a Federal Economic Statistics Advisory Committee session on residual seasonal adjustment. Staff worked with Economic Statistical Methods Division (ESMD) staff to plan a seasonal adjustment workshop.

Staff: Brian Monsell (x31721), Tucker McElroy, James Livsey, Osbert Pang, Anindya Roy, Thomas Trimbur, William R. Bell (ADRM)

B. Seasonal Adjustment Software Development and Evaluation

Description: The goal of this project is a multi-platform computer program for seasonal adjustment, trend estimation, and calendar effect estimation that goes beyond the adjustment capabilities of the Census X-11 and Statistics Canada X-11-ARIMA programs, and provides more effective diagnostics. The goals for FY 2015 include: continuing to develop a version of the X-13ARIMA-SEATS program with accessible output and updated source code so that, when appropriate, the Economic Directorate can produce SEATS adjustments; and incorporating further improvements to the X-13ARIMA-SEATS user interface, output and documentation. In coordination and collaboration with the Time Series and Related Methods Staff of the Economic Statistical Methods Division (ESMD), staff will provide internal and/or external training in the use of X-13ARIMA-SEATS and the associated programs, such as X-13-Graph, when appropriate.

Highlights: Staff continued developing a version of the X-13ARIMA-SEATS software for testing by the Economic Directorate. New features in this program include improved accessible HTML output, two new diagnostics for adequacy of residuals (Friedman's, Durbin Watson), tables for outlier adjusted SEATS seasonal adjustment and irregular, more diagnostics to .udg output, and correct stock length of month regressors. After testing was completed by the Economic Directorate, Version 1.1, Build 26 of X-13ARIMA-SEATS was released to the public.

Staff continued to add diagnostics to working versions of the software, including an option to the spectrum spec to generate quarterly seasonality diagnostics for monthly series and timer information to the diagnostics files with the specification of a runtime argument (-t). Staff revised the code related to user-defined regressors to add user-defined types for trading day, constant, length of month/quarter, and outliers (AO, LS and SO), and checked how multiple types of user-defined regressors worked with the chi-squared and F-statistics.

Staff also fixed defects in the irregular regression modeling routines and the SEATS routines and changed how the F-test for trading day computes the degrees of freedom for the test. Staff improved the output of the regARIMA modeling, sliding spans and revisions history routines for series with large values. Staff continues to develop a version of the X-13ARIMA-SEATS program with updated SEATS routines.

Staff continued the development of sigex, a suite of R routines for modeling multivariate time series. Staff revised the software to include a windowing method to perform the signal extraction and improve modeling procedures.

Staff compiled the source code for the latest version of regCMPNT, fixed defects in the code, and ran examples to test the executable. Staff developed general purpose moving holiday routines in R for high frequency time series for a joint project with the Economic Directorate and Palentir consulting. In the course of this work, staff noted that economic activity was depressed on the day of Easter Sunday. Given those observations, staff developed an Easter[0] regressor into a working version of X-13ARIMA-SEATS to support research into the use of this regressor in monthly retail series published by the Economic Directorate.

Staff: Brian Monsell (x31721), Tucker McElroy, Osbert Pang

C. Research on Seasonal Time Series—Modeling and Adjustment Issues

Description: The main goal of this research is to discover new ways in which time series models can be used to improve seasonal and calendar effect adjustments. An important secondary goal is the development or improvement of modeling and adjustment diagnostics. This fiscal year’s projects include: (1) continuing research on goodness of fit diagnostics (including signal extraction diagnostics and Ljung-Box statistics) to better assess time series models used in seasonal adjustment; (2) studying the effects of model based seasonal adjustment filters; (3) studying multiple testing problems arising from applying several statistics at once; (4) determining if information from the direct seasonally adjusted series of a composite seasonal adjustment can be used to modify the components of an indirect seasonal adjustment, and more generally investigating the topics of benchmarking and reconciliation for multiple time series; (5) studying alternative models of seasonality, such as Bayesian and/or long memory models and/or heteroskedastic models, to determine if improvement to seasonal adjustment methodology can be obtained; (6) studying the modeling of stock holiday and trading day on Census Bureau time series; (7) studying methods of seasonal adjustment when the data are no longer univariate or discrete (e.g., multiple frequencies or multiple series); (8) studying alternative seasonal adjustment methods that may reduce revisions or have alternative properties; and (9) studying nonparametric methods for estimating regression effects and their behavior under long range dependence and/or extreme values.

Highlights: During FY 2016, staff (a) conducted simulation and empirical work to vet new methodologies for fitting vector moving average models; (b) developed estimators for initial values needed to compute signal extraction estimates in a state space framework; (c) modeled weather data from the National Climatic Data Center (obtained through a web scraping tool) to be used in a weather-assisted seasonal adjustment of construction series; and (d) modeled daily time series (New Zealand immigration data, and credit card transaction data) with multiple forms of seasonality, and utilized software from the Bureau of Labor Statistics to obtain seasonal adjustments with new work on the week-to-month reporting problem.

Staff also (a) developed algorithms for quickly computing signal extraction estimates from long samples of high frequency data, and (b) examined mean squared error in simulation of seasonal adjustment from the X11 and SEATS software.

Staff: Tucker McElroy (x33227), James Livsey, Brian Monsell, Osbert Pang, William Bell (ADRM), David Findley (Private Collaborator)

D. Supporting Documentation and Software for X-13ARIMA-SEATS

Description: The purpose of this project is to develop supplementary documentation and utilities for X-13ARIMA-SEATS that enable both inexperienced seasonal adjustors and experts to use the program as effectively as their backgrounds permit. This fiscal year’s goals include improving the X-13ARIMA-SEATS documentation, exploring the use of R packages that interface with X-13ARIMA-SEATS, and exploring the use of component and Java software developed at the National Bank of Belgium.

Highlights: Staff updated the *X-13ARIMA-SEATS REFERENCE MANUAL* to include information on new options and diagnostics and updated the "Getting Started" papers used to introduce users to X-13ARIMA-SEATS. Staff updated HTML files to release new versions of X-13ARIMA-SEATS and Win X-13, and updated HTML documentation files for Win X-13.

Staff worked with Christoph Sax to improve utilities related to the seasonal R package. Staff also collaborated with Christoph Sax in creating an interface to improve the Census Bureau's communication of seasonal adjustment and X-13ARIMA-SEATS.

Staff updated software license and disclaimer statement after consulting with Commerce Department lawyers. Staff worked with CSRM and CSM staff to facilitate the reorganization of the research report series. Staff developed a generalized routine to develop moving holiday regressors for weekly and daily time series. Staff developed a modified Hough transform line detector in Matlab.

Staff: Brian Monsell (x31721), Tucker McElroy, James Livsey, Osbert Pang, Daniel Weinberg

E. Missing Data Adjustment Methods for Product Data in the Economic Census

Description: The Economic Census collects general items from business establishments such as total receipts, as well as more detailed items such as product sales. Although product data are an essential component of the Economic Census, item response rate is low. This project investigates methods for imputing missing product data in the Economic Census. Staff researched three methods for treating missing product line data: expansion estimation, hot deck (random and nearest neighbor), and sequential regression multivariate imputation (SRMI). Staff was asked to apply the SRMI method to these data and assist in making a recommendation.

Highlights: During FY 2016, staff was integrally involved in applying classification trees to determine characteristics of industries for which one variation of hot deck outperformed the other (random hot deck versus nearest neighbor hot deck). Staff worked with researchers from the Economic Directorate on a presentation and paper titled, "Using Classification Trees to Recommend Hot Deck Imputation Methods: A Case Study" for the 2015 FCSM conference. Staff also presented this work internally to the editing and imputation knowledge sharing community group. This project is now complete.

Staff: Darcy Steeg Morris (x33989), Maria Garcia, Yves Thibaudeau

1.14 2012 COMMODITY FLOW SURVEY (Economic Project 7103012)

A. 2012 Commodity Flow Survey

Description: This project provides a retrospective analysis of the cost-quality tradeoffs that the Commodity Flow Survey (CFS) made moving from a 2007 paper-only to a 2012 paper and electronic multi-mode data collection strategy. Based on the data quality findings, the possibility of adding additional edits or modifications to the instruments will be investigated. Optimization strategies for a multi-mode data collection strategy in the 2017 CFS and cost-quality implications of an all-electronic collection will be studied.

Highlights: During FY 2016, staff completed the final research report which summarized the results and recommendations from the project. Among the results, staff found some significant differences between the response modes in terms of data quality and probability of responding to future quarters of the survey. This project is now complete.

Staff: Robert Ashmead (x31564), Eric Slud, Joanna Fane Lineback (CSM)

1.15 INVESTIGATION OF ALTERNATIVE METHODS FOR RESOLVING BALANCE COMPLEX FAILURES IN StEPS (Economic Project TBA)

A. Investigation of Alternative Methods for Resolving Balance Complex Failures in StEPS

Description: The Standard Economic Processing System (StEPS) implements a raking algorithm for adjusting balance complexes in order to satisfy the requirement that the sum of items (details) in a balance complex balances to reported totals. In this project, we research alternative methods to raking when the data items are negative or when there is subtraction in the balance complex.

Highlights: During FY 2016, staff collaborated with Economic Statistical Methods Division (ESMD) staff on research with the goal of developing alternative methods to raking for the Standard Economic Processing System (StEPS). The StEPS generalized system implements a raking algorithm for adjusting balance complexes in order to satisfy the requirement that the sum of details balances to reported totals. The StEPS raking algorithm was developed for positive data only. The raking adjustment fails when the balance complex includes subtraction or if detail items are allowed to take negative values. Staff developed four separate alternative optimization methods based on solving a nonlinear programming problem; the objective function minimizes the change between the final and reported details while the constraints ensure the raked details add up to the total. Staff met, discussed, and learned from subject matter analysts how they resolved (manually) most failing balance complexes. Staff is incorporating their techniques within the computer program implementing the methodology.

Staff: Maria Garcia (x31703), Yves Thibaudeau, Laura Bechtel (ESMD)

1.16 BUSINESS DYNAMICS STATISTICS—EXPORT FILE WEIGHTING ISSUE (Research and Methodology Directorate TBA)

A. Business Dynamics Statistics—Export File Weighting Issue

Description: The challenge: we are unable to match the universe of export transactions to firms on the business register. Therefore, we cannot identify the universe of firms that export U.S. goods. We can pursue two options—(i) produce business dynamics statistics based on the identified cases only. For example, an official Census Bureau data product, the *Profile of U.S. Importing and Exporting Companies*, is released based on the “known” matches and users are provided with a technical documentation explaining the data limitations; or (ii) construct weights to create business dynamics statistics that are representative of the U.S. exporter population.

Highlights: During FY 2016, staff researched the problem of improving matching rates between the Census Bureau’s Foreign Trade Export transactions data (XP) or Innovative firms patents data (IF) to the Business Register (BR). The objective is to add variables from the BR (employment, age, industry, etc.) needed to publish Business Statistics Dynamics (BDS). Staff researched constructing weights using data from other Census Bureau data sources (e.g. Economic Census Data). In discussions with XP, IF, and Economic Census subject matter experts, staff concluded this is not a feasible solution. Staff proposed two alternative approaches. One approach is to calculate a measure of the quality of the existing matches and their representativeness in the XP and IF data files using the R-indicator introduced in Schouten, Cobben and Bethlehem (2009). Alternatively, staff proposed setting up the problem of augmenting the exports and patents datasets with variables from the Business Register (BR) as a missing data problem and implementing two separate imputation methods: Statistical Matching (D’Orazio, 2016) and the multiple imputation procedure Sequential Regression Multivariate Imputation (SRMI, Raghunathan et al., 2001). Staff found multiple difficulties in implementing these methods due to the limited number of common variables needed for matching or for identifying covariates for implementing a sequential regression method.

Staff: Maria Garcia (x31703), Emanuel Ben-David

1.17 ASSESSMENT OF FINANCE METHODOLOGY (Administration and CFO Project TBA)

A. Assessment of Census Bureau’s Finance Methodology for Estimating Accruals

Description: Annually, staff members developed and carried out statistical methodology to validate the Finance Division’s current methodology for estimating Census Bureau’s total 2015 accruals as of September 30, 2015. The total FY 2015 accruals is the total expenses, but have not been paid. Without contacting every contractor, there is no way to know this total with certainty. To estimate this total, the Finance Division multiplies the total value of all contracts and purchase orders that have not been paid as of August 31, 2015 by an average estimate of the ratio of amount paid on contracts/purchases orders to total values of contracts/purchases.

Highlights: During FY 2016 and as in previous years, staff designed, selected a sample, and analyzed financial data with the Finance Division and produced a confidence interval to estimate the total accruals. Details and results are contained in the final August 19, 2016 memo to the Finance Division, “*A Sampling Plan to Assess the Census Bureau’s Method for Estimating FY 2015 Accruals”.* Staff also worked with the Finance Division to consider alternatives to the Finance Division’s current methodology.

Staff: Tommy Wright (x31702), Michael Leibert, Carolina Franco, Robin Guinn (FIN), Quiana Johnson (FIN).

1.18 NATIONAL SURVEY OF DRUG USE & HEALTH
(Census Bureau Project 7236045)

A. National Survey of Drug Use & Health
Description: This project is a feasibility study concerning the extension of the National Survey of Drug Use & Health (NSDUH) to Puerto Rico and other U.S. island areas. Our staff will focus specifically on small area estimation methodology and will determine if and how the island areas can be incorporated into the current NSDUH small area estimation methodology.

Highlights: During FY 2016, staff completed and submitted the final report for the Substance Abuse and Mental Health Services Administration (SAMHSA) detailing the considerations for conducting the National Survey on Drug Use and Health in the U.S. island areas. The project is now complete.

Staff: Robert Ashmead (x31564), Jerry Maples

1.19 PROGRAM DIVISION OVERHEAD
(Census Bureau Project 0331000)

A. Center Leadership and Support
This staff provides ongoing leadership and support for the overall collaborative consulting, research, and administrative operation of the center.

Staff: Tommy Wright (x31702), Lauren Emanuel, Michael Hawkins, Michael Leibert, Erica Magruder, Eric Slud, Kelly Taylor, Bill Winkler

B. Research Computing
Description: This ongoing project is devoted to ensuring that Census Bureau researchers have the computers and software tools they need to develop new statistical methods and analyze Census Bureau data.

Highlights: During FY 2016, the Integrated Research Environment (IRE) team continued to develop the IRE, a shared Linux computing platform that will replace the current "compute clusters": research1, research2, and the RDC cluster. The IRE will provide the logical separation of project data and activities currently provided in the RDC environment without using a separate login for each project. A collection of scripts will enable the user to "change into" a particular project where they will be presented only with the data associated with that project. Testing and integrating those scripts with the job scheduler (PBSPro) is the current focus, as well as integrating the system with the CES management system (CMS) and the Data Management System (DMS). In December 2015, staff discovered that the –S option to qsub, which was intended to specify an alternate shell for a job, would accept an arbitrary script. This provided us with a method for initializing the shell in which the job script runs, which was critical. Progress was made in setting the correct DISPLAY variable in interactive PBS sessions to enable the forwarding of X output. Because the project environment must be initialized with a script prior to job execution, software that spawns "worker" processes either on the same node or on other nodes are of concern. Will these worker processes have the correct environment or not? If not, is there a way to initialize the environment for these workers to obtain the correct results? Staff began exploring this question with one such case– Matlab's Parallel Computing Toolbox and plan to test other commonly used parallel software prior to going into production. The initial migration of the RDC environment to the IRE is expected during FY 2017 followed by the internal clusters (research1 and research2).

Staff: Chad Russell (x33215)

2. RESEARCH

2.1 GENERAL RESEARCH AND SUPPORT (Census Bureau Project 0331000)

2.2 GENERAL RESEARCH (Census Bureau Project 0925000)

Missing Data, Edit, and Imputation

Motivation: Missing data problems are endemic to the conduct of statistical experiments and data collection projects. The instigators almost never observe all the outcomes they had set to record. When dealing with sample surveys or censuses that means individuals or entities in the survey omit to respond, or give only part of the information they are being asked to provide. In addition, the information provided may be logically inconsistent, which is tantamount to missing. To compute official statistics, agencies need to compensate for missing data. Available techniques for compensation include cell adjustments, imputation and editing. All these techniques involve mathematical modeling along with subject matter experience.

Research Problems: Compensating for missing data typically involves explicit or implicit modeling. Explicit methods include Bayesian multiple imputation and propensity score matching. Implicit methods revolve around donor-based techniques such as hot-deck imputation and predictive mean matching. All these techniques are subject to edit rules to ensure the logical consistency of remedial product. Research on integrating together statistical validity and logical requirements into the process of imputing continues to be challenging. Another important problem is that of correctly quantifying the reliability of predictors that have been produced in part through imputation, as their variance can be substantially greater than that computed nominally.

Potential Applications: Research on missing data leads to improved overall data quality and predictors accuracy for any census or sample survey with a substantial frequency of missing data. It also leads to methods to adjust the variance to reflect the additional uncertainty created by the missing data. Given the ever rising cost of conducting censuses and sample surveys, imputation and other missing-data compensation methods may come to replace actual data collection, in the future, in situations where collection is prohibitively expensive.

A. Editing
Description: This project covers development of methods for statistical data editing. Good methods allow us to produce efficient and accurate estimates and higher quality microdata for analyses.

Highlights: The StEPS generalized editing system implements a raking algorithm for adjusting balance edits in order to satisfy the requirement that the sum of detail items balances to final totals. The existing raking algorithm fails when the data items are negative or when there is subtraction in the balance complex. Staff developed alternatives to the StEPS raking algorithm based on solving a nonlinear optimization problem that minimizes several loss functions and resolves failing balance complexes when detail items are allowed to be negative. The method ensures that the raked details add up to the final total.

Staff: Maria Garcia (x31703)

B. Editing and Imputation
Description: Under this project, our staff provides advice, develops computer edit/imputation systems in support of demographic and economic projects, implements prototype production systems, and investigates edit/imputation methods.

Highlights: During FY 2016, staff revised a paper submitted for publication that discusses modeling conditional probability to predict missing variables in subsequent waves of a longitudinal survey. Staff researched the possibility of increasing the accuracy of the predictors on mortgage ownership derived from the American Community Survey (ACS) by extracting and exploiting information from the “CoreLogic” commercial “Big Data” data set. Staff is exploring the possibility of creating “equal-propensity strata” where the value of the mortgage question is homogeneous. This could provide predictive information to improve the accuracy of the mortgage item in ACS.

Staff also made several advances developing methods for handling missing data and/or sparse or fragmented data in surveys and administrative lists. In the context of modeling small cells arising from sparse contingency tables, staff developed methods based on log-linear probabilities. In other efforts, staff showed how to combine small data and large data together to reduce total error of small data while also accounting for bias of administrative data.

Staff: Yves Thibaudeau (x31706), Maria Garcia, Martin Klein, Darcy Steeg Morris, Bill Winkler

Record Linkage

Motivation: Record linkage is intrinsic to efficient, modern survey operations. It is used for unduplicating and updating name and address lists. It is used for applications such as matching and inserting addresses for geocoding, coverage measurement, Primary Selection Algorithm during decennial processing, Business Register unduplication and updating, re-identification experiments verifying the confidentiality of public-use microdata files, and new applications with groups of administrative lists. Significant theoretical and algorithmic progress (Winkler 2004ab, 2006ab, 2008, 2009a; Yancey 2005, 2006, 2007, 2011) demonstrates the potential for this research. For cleaning up administrative records files that need to be linked, theoretical and extreme computational results (Winkler 2010, 2011b) yield methods for editing, missing data and even producing synthetic data with valid analytic properties and reduced/eliminated re-identification risk. Easy means of constructing synthetic make it straightforward to pass files among groups.

Research Problems: The research problems are in three major categories. First, we need to develop effective ways of further automating our major record linkage operations. The software needs improvements for matching large sets of files with hundreds of millions of records against other large sets of files. Second, a key open research question is how to effectively and automatically estimate matching error rates. Third, we need to investigate how to develop effective statistical analysis tools for analyzing data from groups of administrative records when unique identifiers are not available. These methods need to show how to do correct demographic, economic, and statistical analyses in the presence of matching error.

Potential Applications: Presently, the Census Bureau is contemplating or working on many projects involving record linkage. The projects encompass the Demographic, Economic, and Decennial areas.

A. Disclosure Avoidance for Microdata

Description: Our staff investigates methods of microdata masking that preserves analytic properties of public-use microdata and avoid disclosure.

Highlights: During FY 2016, staff reviewed twelve papers on variants of differential privacy and their relationship to Yang, Fienberg, and Rinaldo (2012) and Winkler (2010). Staff e-mailed comments to staff in the Center for Disclosure Avoidance Research (CDAR) on how the methods of modeling/edit/imputation in Winkler (1997, 2003, 2008, 2010) can be used for generating synthetic data with valid analytic properties and very significantly reduced re-identification risk. Staff also e-mailed several papers and an extensive list of references on microdata confidentiality to staff in CDAR. Staff e-mailed information and a paper on microdata confidentiality to staff in the Economic Directorate.

Staff refereed two papers for *Statistical Data Protection 2016*. Staff met with a professor from Cornell Tech to discuss issues related to analytic properties of anonymized data. Staff provided a set of papers (Winkler 1997, 2003, 2008, 2010) about analytic and probabilistic constraints on data that can be used to reduce re-identification risk in public-use microdata. Staff met with one individual in the Center for Disclosure Avoidance (CDAR) to discuss microdata confidentiality issues and computational algorithms. Staff met with the Research and Methodology Associate Director to discuss issues related to microdata confidentiality, particularly theoretical extensions of differential privacy to loglinear models. At the direction of the RM Associate Director, staff put together most of the twenty programs for modeling/edit/imputation/ micro-data-confidentiality that are being researched by various individuals within the Census Bureau, Duke University, and Carnegie Mellon University.

Staff: William Winkler (x34729)

B. Record Linkage and Analytic Uses of Administrative Lists

Description: Under this project, staff will provide advice, develop computer matching systems, and develop and perform analytic methods for adjusting statistical analyses for computer matching error.

Highlights: Staff reviewed papers by Hof and Zwinderman (2012, 2015) and Tancredi and Liseo (2015) that had models for adjusting statistical analyses for linkage error. Staff reviewed and sent comments to the authors of three record linkage papers that had already appeared in journals.

Staff e-mailed comments to staff at the National Agriculture Statistics Service who are building a system for the 2017 Agriculture Census. Staff provided advice and an extensive list of references to staff in the Center for Administrative Records Research and Applications (CARRA) for a proposed record linkage application. Staff spent four days finding record linkage software and non-Title 13 files for a demonstration in a Data Integration class at the University of Maryland, College Park (UMD) after backups on Census Bureau computers were lost after new PC backup software failed. Staff e-mailed the current versions of *BigMatch* that were used for 2010 Decennial Census production to staff in the Decennial Statistical Studies Division (DSSD). Staff also gave a large number of comments regarding machine learning and distributed computing research to staff in DSSD.

Staff circulated a review to staff in CARRA, DSSD, DITD, and CSRM indicating that *BigMatch* continues as the fastest record linkage software in the world. Staff provided very extensive advice to a researcher at the National Institutes of Health on record linkage software and methods. Staff e-mailed information and comments related to cleaning up and analyzing national files to staff in the Computer Services Division (CSVD). Staff e-mailed a number of comments on edit/imputation systems to individuals in the CEDCAP project on edit/imputation. Staff did extensive reading of background literature on record linkage.

Staff began comparisons of existing theoretical methodologies for regression analysis of linked data. The main goal of this comparison is to take an approach that can improve upon these methods.

Staff e-mailed considerable information on specific record linkage methods to a staff member at the National Cancer Institute. Staff agreed to be on a Ph.D. committee at the University of Maryland. Staff also e-mailed considerable information related to record linkage to a member of the Federal Reserve Board. While attending the week-long workshop and presenting one of the main technical talks at the Isaac Newton Institute at Cambridge University, staff received strong confirmation that *BigMatch* remains the fastest record linkage software in the world.

Staff: William Winkler (x34729), Ned Porter, Emanuel Ben-David

C. Modeling, Analysis, and Quality of Data
Description: Our staff investigates methods of the quality of microdata primarily via modeling methods and new software techniques that accurately describe one or two of the analytic properties of the microdata.

Highlights: During FY 2016, staff provided comments, advice, a list of references, and a list of shareware/freeware to the New York City government. Staff made comments to DSSD staff related to a research proposal from Carnegie Mellon University. Staff provided extensive comments to staff of the Office of National Statistics in the UK related to the EM algorithm for parameter estimation in record linkage.

Staff completed the first draft software version of set covering algorithms edit generation for a program. The theory is based on Fellegi and Holt (*JASA* 1976), Garfinkel, Kunnathur, and Liepins (*Operations Research* 1986), and Winkler (1997). The first version works on the example of Garfinkel et al. Staff are working on a larger example for the Italian Labour Force Survey using data provided by IBM and ISTAT.

Staff worked on the following problem posed by a colleague: Given two real sequences $y_1 < y_2 < ... < y_N$ and $z_1 < z_2 < ... < z_n$, where $n < N$. We want to find a sequence $x_1 < x_2 < ... < x_n$ to minimize the value: $(z_1 - x_1)^2 + ... + (z_n - x_n)^2$ where x_i is in $\{y_1, ..., y_N\}$. Staff provided a heuristic solution using divide and conquer strategy. Based on a recommendation from another colleague, staff read a few chapters of *Numerical Optimization* about quadratic programming.

Staff provided extensive background on modeling/edit/imputation to individuals working on the CEDCAP edit/imputation project to develop generalized methods/software for the Decennial Census and, possibly, approximately 140 other Census Bureau sample surveys. The background document included a description of the specific work successfully performed at five statistical agencies that have been able to develop Fellegi-Holt systems. The background covered some of the specifics of the computational algorithms and gave a number of references in refereed journals and in agency technical reports. Staff also provided a document on how to develop teams with the technical skills for generalized systems based on successful projects at Statistics Canada and the Census Bureau: Winkler, W. E. and Hidiroglou, M. (1998), “Developing Analytic Programming Ability to Empower the Survey Organization,” http://www.census.gov/srd/papers/pdf/rr9804.pdf.

Staff worked on extending the algorithms for generating implicit edits. The extension uses heuristics that keep track of the most active fields used in generating new edits.

Staff: William Winkler (x34729), Xiaoyun Lu, Ned Porter, Emanuel Ben-David, Maria Garcia

D. R Users Group
Description: The initial objective of the R Users Group is to identify the areas of the Census Bureau where R software is developed and those other areas that could benefit from such development. The scope of the topics is broad and it includes estimation, missing data methods, statistical modeling, Monte-Carlo and resampling methods. The ultimate goal is to move toward integrated R tools for statistical functionality at the Census Bureau.

Initially, the group will review basic skills in R and provide remedial instruction as needed. The first topic for deeper investigation is complex-survey infrastructure utilities, in particular an evaluation of the “Survey” package and its relevance at the Census Bureau in the context of weighing, replication, variance estimation and other structural issues.

Highlights: During FY 2016, staff continued to support R users in CSRM and across the Census Bureau. CSRM provides the main infrastructure for R usage until the IT directorate makes an enterprise solution available.

Staff: Yves Thibaudeau (x31706), Chad Russell

Small Area Estimation

Motivation: Small area estimation is important in light of a continual demand by data users for finer geographic detail of published statistics. Traditional demographic surveys designed for national estimates do not provide large enough samples to produce reliable direct estimates for small areas such as counties and even most states. The use of valid statistical models can provide small area estimates with greater precision, however bias due to an incorrect model or failure to account for informative sampling can result. Methods will be investigated to provide estimates for geographic areas or subpopulations when sample sizes from these domains are inadequate.

Research Problems:

- Development/evaluation of multilevel random effects models for capture/recapture models.
- Development of small area models to assess bias in synthetic estimates.
- Development of expertise using nonparametric modeling methods as an adjunct to small area estimation models.
- Development/evaluation of Bayesian methods to combine multiple models.
- Development of models to improve design-based sampling variance estimates.
- Extension of current univariate small-area models to handle multivariate outcomes.

Potential Applications:

- Development/evaluation of binary, random effects models for small area estimation, in the presence of informative sampling, cuts across many small area issues at the Census Bureau.
- Using nonparametric techniques may help determine fixed effects and ascertain distributional form for random effects.
- Improving the estimated design-based sampling variance estimates leads to better small area models which assumes these sampling error variances are known.
- For practical reasons, separate models are often developed for counties, states, etc. There is a need to coordinate the resulting estimates so smaller levels sum up to larger ones in a way that correctly accounts for accuracy.
- Extension of small area models to estimators of design- base variance.

A. Small Area Methods with Misspecification
Description: In this project, we undertake research on area-level methods with misspecified models, primarily directed at development of diagnostics for misspecification using robust sandwich-formula variances, cross-validation, and others, and on Bayesian estimation of model parameters within two-component Fay-Herriot models.

Highlights: During FY 2016, no progress was made due to departure of key staff in the Center for Disclosure Avoidance Research Division (CDAR). Efforts are currently underway to obtain proper staffing on this project.

Staff: Jerry Maples (x32873), Gauri Datta, Eric Slud

B. Coverage Properties of Confidence Intervals for Proportions in Complex Surveys
Description: This is primarily a simulation project to investigate the coverage behavior of confidence intervals for proportions estimated in complex surveys. The goal is ultimately to inform recommendations for interval estimates in the American Community Survey (ACS), so the issues of main interest are:
(i) whether the current Wald-type intervals (defined as a point-estimator plus or minus a margin-or-error (MOE) estimate) can be improved by empirical-Bayes modifications or by modified forms of intervals known to perform well in the setting of binomial proportion-estimators, (ii) whether failures of coverage in a simulated complex survey can be ascribed to poor estimation of effective sample size or to other aspects of inhomogeneity and clustering in proportions within realistically complex populations, and (iii) whether particular problems arising with coverage of intervals for small proportions can be overcome. Future research might address whether the confidence interval methods developed for single-domain design-based estimates can also be adapted to small area estimates that borrow strength across domains.

Highlights: During FY 2016, staff expanded the simulation study to include scenarios with more clustering and a higher intra-cluster correlation (ICC) in response to a recent related paper by Dean and Pagano (2015). Staff also incorporated two modifications to the effective sample size available in the literature, one proposed by Korn and Graubard (1998) and another by Dean and Pagano (2015). Staff began analysis of results under this more comprehensive simulation design. Staff studied additional questions such as the effect of the ICC on the coverage and length of all intervals, the effect of having larger cluster sizes, and the interaction of both effects. Staff derived two new methods of computing the design effect, incorporated them in the simulation study, and began evaluating the results.

Staff also explored the conjecture that the failure of methods to account for the uncertainty in estimation of the design effect is a main reason for tendency of undercoverage for confidence intervals for proportions in complex surveys. Staff found empirical evidence supporting this conjecture through the simulation study and devised ideas on how to exploit this finding.

Staff gave a presentation at the Federal Conference on Statistical Methodology (FCSM) discussing these results.

Staff: Carolina Franco (x39959), Eric Slud, Thomas Louis (ADRM), Rod Little (University of Michigan)

C. Small Area Estimates of Disability
Description: This project is from the Development Case proposal to create subnational estimates of specific disability characteristics (e.g., number of people with autism). This detailed data is collected in a supplement of the Survey of Income and Program Participation (SIPP). However, the SIPP is only designed for national level estimates. This project is to explore small area models to combine SIPP with the large sample size of the American Community Survey to produce state and county level estimates of reasonable quality.

Highlights: During FY 2016, staff rewrote the codebase to have more flexibility in modeling options. Staff prepared and submitted a manuscript to the *Small Area Special Edition of the Journal of the Royal Statistical Society, Series A*. The paper was reviewed and allowed to submit a revision. Staff is currently revising the manuscript and addressing the referees' comments.

Staff: Jerry Maples (x32873), Amy Steinweg (SEHSD)

D. Using ACS Estimates to Improve Estimates from Smaller Surveys via Bivariate Small Area Estimation Models
Description: Staff will investigate the use of bivariate area-level models to improve small area estimates from one survey by borrowing strength from related estimates from a larger survey. In particular, staff will explore the potential of borrowing strength from estimates from the American Community Survey, the largest U.S. household survey, to improve estimates from smaller U.S. surveys, such as the National Health Interview Survey, the Survey of Income and Program Participation, and the Current Population Survey.

Highlights: During FY 2016, staff prepared and delivered a presentation on related research results at the 2016 Ross-Royall Symposium at Johns Hopkins University.

Staff: Carolina Franco (x39959), William R. Bell (ADRM)

E. Multivariate Fay-Harriot Hierarchical Bayesian Estimation of Small Area Means under Functional Measurement Error
Description: Area-level models have been extensively used in small area estimation to produce model-based estimates of a population characteristic for small areas (e.g., Fay and Herriot, 1979). Multivariate area level models have also been used to jointly model multiple characteristics of correlated responses (e.g., Huang and Bell, 2012, Franco and Bell, 2015). Such models may lead to more precise small area estimates than separate univariate modeling of each characteristic. Typically both univariate and multivariate small area estimation models use auxiliary information to borrow strength from other areas and covariates associated with a response variable or a response vector. However, auxiliary variables are sometimes measured or obtained from sample surveys and are subject to measurement or sampling error. Researchers recognized that ignoring measurement error in the covariates and using standard solutions developed for covariates measured without error may lead to suboptimal inference. It was demonstrated in the univariate small area estimation setup that this naïve approach can result in model-based small area estimators that are more variable than the direct estimators when some of the covariate values in a small area are measured with substantial error (cf. Ybarra and Lohr, 2008, *Biometrika*; Arima, Datta and Liseo, 2015, *Scandinavian Journal of Statistics*). We are investigating a multivariate Fay-Herriot model and develop Bayes small area estimates when one or more auxiliary variables are measured with error. We work out a hierarchical Bayesian analysis for the multivariate Fay-Herriot model with a functional measurement error treatment for the covariates measured with error.

Highlights: During FY 2016, staff investigated the performance of multivariate Fay-Herriot measurement error models with a Bayesian implementation by applying them to estimate yearly changes of poverty rates for school-aged children in counties. Staff initially used the software JAGS for the implementation. Staff empirically compared a bivariate measurement error model to a naïve model where the covariate measured with error is treated as a known covariate.

Staff developed an algorithm specifically tailored to the class of priors for which staff had previously proved the propriety of the posterior under some conditions. The algorithm uses a combination of Metropolis-Hastings and Gibbs sampling and can be applied to models of high dimensions as well as to large data sets. Staff wrote a program to implement this algorithm and began debugging it. Though staff initially implemented a bivariate Fay-Herriot Model with Measurement Error using JAGS, it is unclear whether it is possible to use JAGS to implement this model for problems with higher dimensions. It is also not straightforward to use JAGS to implement the class of priors for which theory was developed. Moreover, staff discovered that plain Gibbs sampling is inefficient for this type of model and other similar linear models. In fact, Gibbs sampling is so slow that it is not feasible to apply it to large data sets (i.e., county-level school-aged children in poverty) even for relatively low dimensions (i.e., bivariate models). This motivated staff to develop a specialized algorithm. Staff collected the empirical, theoretical, and computational results and completed a paper that is currently under revision in a peer reviewed publication.

Staff also studied the theoretical and practical differences between functional and structural measurement error models, both analytically and via simulation studies.

Staff presented some of the results in two invited talks: the 8th International Conference of the ERCIM Working Group on Computational and Methodological Statistics (CMStatistics 2015) at the University of London, United Kingdom and the Small Area Estimation Conference 2016 in Maastricht, Netherlands.

Staff: Carolina Franco (x39959), Gauri Datta, William R. Bell (ADRM)

F. **Smoothing Design Effects for Small Sample Areas**
Description: In Small Area Estimation, the design-based estimates for many areas are based on very small samples. We propose using information from a larger aggregate, whose design-based variance estimator can be reliably estimated to inform us about the design effect at the small component area. Our goal is to create a principled method to use information about design effects at the higher level to estimate design effects at the lower level. Due to the lack of data, this will require strong assumptions and large amounts of smoothing of design features over the small local areas.

Highlights: During FY 2016, staff created a framework based on a pseudo stratified design to link the design effects at the local area to the variance estimate of the larger aggregate area. In order to preserve as much design information about the local area as possible, effects for unequal sample size, clustering structure and informative sampling (survey outcome related to probability of selection) are first conditioned out so that the residual design effect is what will be estimated from the higher level aggregate. In applications where these design effects are used to smooth out the design-based sampling variance estimates, procedures were developed when there was a model for the underlying rate (with associated covariate predictors) and when there was no underlying model (no covariate predictors). When there were no predictors, an optimization criterion was used to determine the best-weighted average of the local area level estimated rate and the group-level estimated rate. Results from research were presented in a talk at the Joint Statistical Meetings in Chicago and a report was submitted for the proceedings of the conference.

Staff: Jerry Maples (x32873)

Survey Sampling-Estimation and Modeling

Motivation: The demographic sample surveys of the Census Bureau cover a wide range of topics but use similar statistical methods to calculate estimation weights. It is desirable to carry out a continuing program of research to improve the accuracy and efficiency of the estimates of characteristics of persons and households. Among the methods of interest are sample designs, adjustments for non-response, proper use of population estimates as weighting controls, small area estimation, and the effects of imputation on variances.

The Economic Directorate of the Census Bureau encounters a number of issues in sampling and estimation in which changes might increase the accuracy or efficiency of the survey estimates. These include, but are not restricted to, a) estimates of low-valued exports and imports not currently reported, b) influential values in retail trade survey, and c) surveys of government employment.

The Decennial Census is such a massive undertaking that careful planning requires testing proposed methodologies to achieve the best practical design possible. Also, the U.S. Census occurs only every ten years and is the optimal opportunity to conduct evaluations and experiments with methodologies that might improve the next census. Sampling and estimation are necessary components of the census testing, evaluations, and experiments. The scale and variety of census operations require an ongoing research program to achieve improvements in methodologies. Among the methods of interest are coverage measurement sampling and estimation, coverage measurement evaluation, evaluation of census operations, uses of administrative records in census operations, improvements in census processing, and analyses that aid in increasing census response.

Research Problems:

• How can methods making additional use of administrative records, such as model-assisted and balanced sampling, be used to increase the efficiency of household surveys?

• Can non-traditional design methods such as adaptive sampling be used to improve estimation for rare characteristics and populations?

• How can time series and spatial methods be used to improve ACS estimates or explain patterns in the data?

- Can generalized weighting methods be implemented via optimization procedures that allow better understanding of how the various steps relate to each other?
- Some unusual outlying responses in the surveys of retail trade and government employment are confirmed to be accurate, but can have an undesired large effect on the estimates - especially estimates of change. Procedures for detecting and addressing these influential values are being extended and examined through simulation to measure their effect on the estimates, and to determine how any such adjustment best conforms with the overall system of estimation (monthly and annual) and benchmarking.
- What models aid in assessing the combined effect of all the sources of estimable sampling and nonsampling error on the estimates of population size?
- How can administrative records improve census coverage measurement, and how can census coverage measurement data improve applications of administrative records?
- What analyses will inform the development of census communications to encourage census response?
- How should a national computer matching system for the Decennial Census be designed in order to find the best balance between the conflicting goals of maximizing the detection of true duplicates and minimizing coincidental matches? How does the balance between these goals shift when modifying the system for use in other applications?
- What can we say about the additional information that could have been obtained if deleted census persons and housing units had been part of the Census Coverage Measurement (CCM) Survey?

Potential Applications:

- Improve estimates and reduce costs for household surveys via the introduction of additional design and estimation procedures.
- Produce improved ACS small area estimates through the use of time series and spatial methods.
- Apply the same weighting software to various surveys.
- New procedures for identifying and addressing influential values in the monthly trade surveys could provide statistical support for making changes to weights or reported values that produce more accurate estimates of month-to-month change and monthly level. The same is true for influential values in surveys of government employment.
- Provide a synthesis of the effect of nonsampling errors on estimates of net census coverage error, erroneous enumerations, and omissions and identify the types of nonsampling errors that have the greatest effects.
- Describe the uncertainty in estimates of foreign-born immigration based on American Community Survey (ACS) used by Demographic Analysis (DA) and the Postcensal Estimates Program (PEP) to form estimates of population size.
- Improve the estimates of census coverage error.
- Improve the mail response rate in censuses and thereby reduce the cost.
- Help reduce census errors by aiding in the detection and removal of census duplicates.
- Provide information useful for the evaluation of census quality.
- Provide a computer matching system that can be used with appropriate modifications for both the Decennial Census and several Decennial-related evaluations.

A. Household Survey Design and Estimation

[See Demographic Projects]

B. Sampling and Estimation Methodology: Economic Surveys

Description: The Economic Directorate of the Census Bureau encounters a number of issues in sampling and estimation in which changes might increase the accuracy or efficiency of the survey estimates. These include estimates of low-valued exports not currently reported, alternative estimation for the *Quarterly Financial Report*, and procedures to address nonresponse and reduce respondent burden in the surveys. Further, general simulation software might be created and structured to eliminate various individual research efforts. An observation is considered influential if the estimate of total monthly revenue is dominated by its weighted contribution. The goal of the research is to find methodology that uses the observation but in a manner that assures its contribution does not dominate the estimated total or the estimates of period-to-period change.

Highlights: Staff collaborated with staff in the Economic Directorate to find an innovative solution to the basic and previously unanswered question of how to develop initial settings for the parameters required to implement M-estimation methodology for detecting and treating verified influential values in economic surveys. The economic populations of interest are highly skewed and are consequently highly stratified, making normal distribution theory inapplicable. In addition, the Census Bureau conducts a large number of economic surveys monthly and must publish results quickly. While the team investigated settings for several parameters, the most challenging problem was to develop an "automatic" data-driven method for setting the initial value of the tuning constant φ, the parameter with the greatest influence on performance of the algorithm. Of all the methods considered, the team found that the methods defined in terms of the requirement for the accuracy of published estimates, namely coefficients of variation and standard errors, yielded the best performance when judged in terms of lack of convergence issues for the algorithm and appropriate detections.

In addition, these methods can be implemented on a large scale for a wide variety of population distributions. The methodology and an empirical analysis of 36 consecutive months of data from 19 industries in the Monthly Wholesale Trade Survey (MWTS) was presented in an invited paper at Fifth International Conference on Establishment Surveys (ICES-V). Preparations are underway to run a side-by-side test of the methodology in the near future so that MWTS staff can evaluate the effectiveness of the methodology when using it in a production setting.

The team also continued working on a research note concerning the method of Clark Winsorization, an alternative method for detecting and treating influential values that the team investigated before deciding to pursue the M-estimation method. In the research note, the team presents the insights gained about the performance of Clark Winsorization in detecting influential values.

Staff: Mary Mulry (x31759)

C. The Ranking Project: Methodology Development and Evaluation

Description: This project undertakes research into the development and evaluation of statistical procedures for using sample survey data to rank several populations with respect to a characteristic of interest. The research includes an investigation of methods for quantifying and presenting the uncertainty in an estimated ranking of populations. As an example, a series of ranking tables are released from the American Community Survey in which the fifty states and the District of Columbia are ordered based on estimates of certain characteristics of interest.

Highlights: During FY 2016, staff worked on drafts of three visualizations of rankings using 2011 American Community Survey (ACS) “Travel Time to Work” data. One of the visualizations provides comparisons of pairs of means for the 51 (including Washington, D.C.) states; the second shows a bootstrap distribution for the estimated rank of each state; and the third visualization shows the bootstrap estimates of probability that the estimated rank for state *i* exceeds the estimated rank for state *j*. Preparation for an internet website began. Staff obtained ten years of ACS data on more than 85 variables for ranking; and continued to experiment with various visualizations. Staff worked on a draft paper with supporting theory for ranking methodology and associated uncertainty.

Staff: Tommy Wright (x31702), Martin Klein, Jerzy Wieczorek (Carnegie Mellon University), Brett Moran, Nathan Yau, Michael Leibert

D. Sampling and Apportionment

Description: This short-term effort demonstrated the equivalence of two well-known problems–the optimal allocation of the fixed overall sample size among L strata under stratified random sampling and the optimal allocation of the H = 435 seats among the 50 states for the apportionment of the U.S. House of Representatives following each decennial census. This project continues development with new sample allocation algorithms.

Highlights: Staff published a research report detailing exact optimal allocation algorithms given stated desired precision and given fixed budget. The algorithms all follow from a simple decomposition of sampling error. Optimality of an algorithm with budget constraints is proven.

Staff: Tommy Wright (x31702), Andrew Perry, Adam Maidman

E. Analysis and Estimation of Daily Response Propensities and Use of Contact History Instrument (CHI)

Description: Staff will conduct general research methodology to work on existing files to improve modeling accuracy and to provide suggestions for developing and using response propensities. To help the research, we make use of National Crime Victimization Survey data. Staff is also currently using general research methodology to work on simulation study to describe how to reproduce generalized boosted regression modeling algorithm for estimating propensity scores with bootstrap and continuous treatment methods.

Highlights: During FY 2016, staff developed methods to fit and evaluate models that can predict daily response propensities. Staff updated existing methodology to describe how to fit daily response propensities along with actual survey indicators and survey outcomes to (1) evaluate model accuracy and determine whether the models need refinement; (2) investigate relationship between response propensity and key survey variables; and (3) determine how the daily response propensities may be used to manage fieldwork.

A formal report of the analysis and estimation documenting work on NCVS daily response propensity modeling and methodology for FY 2016 is completed and the final report is currently under review. Staff plan to show how this response propensity methodology can provide potential intervention strategy for survey efforts to increase response rates and how more interview cases can result in completions.

Staff: Isaac Dompreh (x36801), Joseph Schafer (ADRM)

Time Series and Seasonal Adjustment

Motivation: Seasonal adjustment is vital to the effective presentation of data collected from monthly and quarterly economic surveys by the Census Bureau and by other statistical agencies around the world. As the developer of the X-13ARIMA-SEATS Seasonal Adjustment Program, which has become a world standard, it is important for the Census Bureau to maintain an ongoing program of research related to seasonal adjustment methods and diagnostics, in order to keep X-13ARIMA-SEATS up-to-date and to improve how seasonal adjustment is done at the Census Bureau.

Research Problems:

- All contemporary seasonal adjustment programs of interest depend heavily on time series models for trading day and calendar effect estimation, for modeling abrupt changes in the trend, for providing required forecasts, and, in some cases, for the seasonal adjustment calculations. Better methods are needed for automatic model selection, for detection of inadequate models, and for assessing the uncertainty in modeling results due to model selection, outlier identification and non-normality. Also, new models are needed for complex holiday and calendar effects.
- Better diagnostics and measures of estimation and adjustment quality are needed, especially for model-based seasonal adjustment.
- For the seasonal, trading day and holiday adjustment of short time series, meaning series of length five years or less, more research into the properties of methods usually used for longer series, and perhaps into new methods, are needed.

Potential Applications:

- To the effective presentation of data collected from monthly and quarterly economic surveys by the Census Bureau and by other statistical agencies around the world.

A. Seasonal Adjustment

Description: This research is concerned with improvements to the general understanding of seasonal adjustment and signal extraction, with the goal of maintaining, expanding, and nurturing expertise in this topic at the Census Bureau.

Highlights: During FY 2016, staff (a) completed a draft on a maximum entropy extreme value adjustment of New Zealand agricultural time series; (b) continued empirical work on seasonal heteroscedasticity models to show improved forecasting and seasonal adjustment of construction series; (c) extended work on signal extraction decompositions allowing for correlation between components; and (d) worked on a simulation study comparing mean squared errors for seasonal adjustments (and trends) using X11 and SEATS.

Staff: Tucker McElroy (x33227), James Livsey, Brian Monsell, Osbert Pang, Anindya Roy

B. Time Series Analysis

Description: This research is concerned with broad contributions to the theory and understanding of discrete and continuous time series, for univariate or multivariate time series. The goal is to maintain and expand expertise in this topic at the Census Bureau.

Highlights: During FY 2016, staff (a) continued work on stable parametrizations of VARMA models fitted under parameter constraints, utilizing a LASSO objective function; (b) continued simulation and software development for multivariate count time series; (c) further developed and tested likelihood ratio tests for Granger non-causality, as a way to exclude extraneous data from multivariate forecasting problems; (d) developed Bayesian framework for obtaining signal estimates from combined public and private information sources; (e) completed a corrigendum, extending subsampling results for Lipschitz continuous statistics; (f) completed extensive revisions and simulations for work on non-nested model comparisons; (g) developed software to compute autocorrelations for a spatial long memory process; (h) completed study of the multivariate bullwhip effect for retail supply chains; (i) continued research and simulations for two tests of co-integration, one based upon fitted structural models and another based on nonparametric spectral estimates; (j) continued research on co-integration tests, giving a seminar and conducting simulation studies; (k) applied method of computing residual entropy to data sets, and began writing draft paper; (l) further developed software and methods for space-time signal extraction, with a Bayesian component; (m) continued work on Frobenius norm tool, developing more application of the so-called method-of-moments estimators; and (n) continued implementation of vector band pass filters, utilizing canonical trends and cycles to get adequate modeling of series.

Staff: Tucker McElroy (x33227), David Findley (Private Collaborator), Brian Monsell, James Livsey, Osbert Pang, Anindya Roy

C. Time Series Model Development

Description: This work develops a flexible integer-valued autoregressive (AR) model for count data that contain data over- or under-dispersion (i.e. count data where the variance is larger or smaller than the mean, respectively). Such a model will contain Poisson and Bernoulli AR models as special cases.

Highlights: During FY 2016, staff continued to develop theoretical results and computational codes in R to analyze relevant data.

Staff: Kimberly Sellers (x39808)

Experimentation and Statistical Modeling

Motivation: Experiments at the Census Bureau are used to answer many research questions, especially those related to testing, evaluating, and advancing survey sampling methods. A properly designed experiment provides a valid, cost-effective framework that ensures the right type of data is collected as well as sufficient sample sizes and power are attained to address the questions of interest. The use of valid statistical models is vital to both the analysis of results from designed experiments and in characterizing relationships between variables in the vast data sources available to the Census Bureau. Statistical modeling is an essential component for wisely integrating data from previous sources (e.g., censuses, sample surveys, and administrative records) in order to maximize the information that they can provide.

Research Problems:

- Investigate bootstrap methodology for sample surveys; implement the bootstrap under complex sample survey designs; investigate variance estimation for linear and non-linear statistics and confidence interval computation; incorporate survey weights in the bootstrap; investigate imputation and the bootstrap under various non-response mechanisms.
- Investigate methodology for experimental designs embedded in sample surveys; investigation of large-scale field experiments embedded in ongoing surveys; design based and model based analysis and variance estimation incorporating the sampling design and the experimental design; factorial designs embedded in sample surveys and the estimation of interactions; testing non-response using embedded experiments. Use simulation studies.
- Assess feasibility of established design methods (e.g., factorial designs) in Census Bureau experimental tests.
- Identify and develop statistical models (e.g., loglinear models, mixture models, and mixed-effects models) to characterize relationships between variables measured in censuses, sample surveys, and administrative records.
- Assess the applicability of post hoc methods (e.g., multiple comparisons and tolerance intervals) with future designed experiments and when reviewing previous data analyses.

Potential Applications:

- Modeling approaches with administrative records can help enhance the information obtained from various sample surveys.
- Experimental design can help guide and validate testing procedures proposed for the 2020 Census.
- Expanding the collection of experimental design procedures currently utilized with the American Community Survey.

A. Design and Analysis of Embedded Experiments

Description: This ongoing project will explore rigorous analysis of embedded experiments: from simple idealized designs to complex designs used in practice at the Census Bureau.

Highlights: Staff investigated variance estimation in the setting where two experimental treatments are compared in a sample collected from a finite population. A naive bootstrap estimator was seen to be nearly equivalent to an estimator proposed by Van den Brakel & Renssen (1996). Staff also evaluated more sophisticated bootstrap and permutation methods using several sampling designs. The Van den Brakel & Renssen estimator was found to have the smallest bias, though other methods may yield smaller variability under complex sampling designs. Staff began preparing a review of embedded experiments methodology with a view toward applications at the Census Bureau.

Staff: Thomas Mathew (x35337), Andrew Raim, Robert Ashmead

B. Multivariate Nonparametric Tolerance Regions

Description: A tolerance region for a multivariate population is a region computed using a random sample that will contain a specified proportion or more of the population, with a given confidence level. Typically, tolerance regions that have been computed for multivariate populations are elliptical in shape. A difficulty with an elliptical region is that it cannot provide information on the individual components of the measurement vector. However, such information can be obtained if we compute tolerance regions that are rectangular in shape. This project applies bootstrap ideas to compute multivariate tolerance regions in a nonparametric framework. Such an approach can be applied to multivariate economic data and aid in the editing process by identifying multivariate observations that are outlying in one or more attributes and subsequently should undergo further review.

Highlights: During FY 2016, significant progress has been made on developing the necessary theoretical framework. The approach consists of trimming the multivariate data set by employing statistical data depth, and utilizing the extremes of the trimmed dataset as the faces of the hyper-rectangular region. A strategy for determining the number of points to be trimmed is developed, and an algorithm is provided for implementing the methodology. An extensive coverage study shows the favorable performance of the algorithm for moderate to large sample sizes. For smaller sample sizes, a bootstrap calibration routine is recommended for improved performance. A manuscript based on the work is under preparation.

Staff: Thomas Mathew (x35337)

C. Master Address File (MAF) Research—Developing a Generalized Regression Model for Count Data
Description: This project develops a zero-inflated version of a generalized regression model for count data based on the Conway-Maxwell-Poisson distribution to allow for data-dispersion and excess zeroes in the dataset. The objective of this project is to develop and consider an alternative regression model for use to describe associations with changes in the number of housing units (adds or deletes) on a block, and predict where housing growth or decline may occur in the MAF.

Highlights: During FY 2016, staff published a manuscript describing the statistical methodology associated with this work in *Computational Statistics & Data Analysis.* This project is now complete.

Staff: Kimberly Sellers (x39808), Andrew Raim

D. Development of a Bivariate Distribution for Count Data where Data Dispersion is Present
Description: This project develops a bivariate form of the Conway-Maxwell-Poisson distribution to serve as a tool to describe variation and association for two count variables that express over- or under-dispersion (relationships where the variance of the data is larger or smaller than the mean, respectively).

Highlights: During FY 2016, staff published a manuscript associated with this work in the *Journal of Multivariate Analysis.*

Staff: Kimberly Sellers (x39808), Darcy Steeg Morris

E. Developing a Flexible Stochastic Process for Significantly Dispersed Count Data
Description: The Bernoulli and Poisson are two popular count processes; however, both rely on strict assumptions that motivate their use. CSRM staff (with other collaborators) instead propose a generalized count process (hereafter named the Conway-Maxwell-Poisson process) that not only includes the Bernoulli and Poisson processes as special cases, but also serves as a flexible mechanism to describe count processes that approximate data with over- or under-dispersion. Staff introduce the process and its associated generalized waiting time distribution with several real-data applications to illustrate its flexibility for a variety of data structures. This new generalized process will enable analysts to better model count processes where data dispersion exists in a more accommodating and flexible manner.

Highlights: During FY 2016, staff had a manuscript from this work accepted for publication in *The American Statistician*.

Staff: Kimberly Sellers (x39808), Darcy Steeg Morris

F. Analysis of Under-dispersed Count Data
Description: This research concerns contributions to the theory and understanding of under-dispersed count data, and models that accommodate such data. The goal is to expand understanding and expertise in this area at the Census Bureau.

Highlights: During FY 2016, staff wrote a literature review of the topic, addressing causes of data under-dispersion and noting various statistical models that accommodate either only under-dispersion, or over- or under-dispersion. Staff submitted the manuscript for review with a journal and also presented initial research and results at an invited session at the XXVIIIth International Biometric Conference in Victoria, BC, Canada.

Staff: Kimberly Sellers (x39808), Darcy Morris

Simulation and Statistical Modeling

Motivation: Simulation studies that are carefully designed under realistic survey conditions can be used to evaluate the quality of new statistical methodology for Census Bureau data. Furthermore, new computationally intensive statistical methodology is often beneficial because it can require less strict assumptions, offer more flexibility in sampling or modeling, accommodate complex features in the data, enable valid inference where other methods might fail, etc. Statistical modeling is at the core of the design of realistic simulation studies and the development of intensive computational statistical methods. Modeling also enables one to efficiently use all available information when producing estimates. Such studies can benefit from software such as *Tea* for data processing. Statistical disclosure avoidance methods are also developed and properties studied.

Research Problems:

- Systematically develop an environment for simulating complex surveys that can be used as a test-bed for new data analysis methods.
- Develop flexible model-based estimation methods for survey data.
- Develop new methods for statistical disclosure control that simultaneously protect confidential data from disclosure while enabling valid inferences to be drawn on relevant population parameters.
- Investigate the bootstrap for analyzing data from complex sample surveys.
- Continue to formalize the codebase and user interfacing for *Tea*, especially within the context of the current enterprise environment.
- Develop models for the analysis of measurement

errors in Demographic sample surveys (e.g., Current Population Survey or the Survey of Income and Program Participation).

- Identify and develop statistical models (e.g., loglinear models, mixture models, and mixed-effects models) to characterize relationships between variables measured in censuses, sample surveys, and administrative records.
- Investigate noise multiplication for statistical disclosure control.

Potential Applications:

- Simulating data collection operations using Monte Carlo techniques can help the Census Bureau make more efficient changes.
- Use noise multiplication or synthetic data as an alternative to top coding for statistical disclosure control in publicly released data. Both noise multiplication and synthetic data have the potential to preserve more information in the released data over top coding.
- Rigorous statistical disclosure control methods allow for the release of new microdata products.
- *Tea* provides modeling and editing flexibility, especially with a focus on incorporating administrative data.
- Using an environment for simulating complex surveys, statistical properties of new methods for missing data imputation, model-based estimation, small area estimation, etc. can be evaluated.
- Model-based estimation procedures enable efficient use of auxiliary information (for example, Economic Census information in business surveys), and can be applied in situations where variables are highly skewed and sample sizes are not sufficiently large to justify normal approximations. These methods may also be applicable to analyze data arising from a mechanism other than random sampling.
- Variance estimates and confidence intervals in complex surveys can be obtained via the bootstrap.
- Modeling approaches with administrative records can help enhance the information obtained from various sample surveys.

A. Development and Evaluation of Methodology for Statistical Disclosure Control

Description: When survey organizations release data to the public, a major concern is the protection of individual records from disclosure while maintaining quality and utility of the released data. Procedures that deliberately alter data prior to their release fall under the general heading of statistical disclosure control. This project develops new methodology for statistical disclosure control, and evaluates properties of new and existing methods. We develop and study methods that yield valid statistical analyses, while simultaneously protecting individual records from disclosure.

Highlights: During FY 2016, staff continued work on the development of new finite sample methods for drawing parametric inference based on singly imputed partially synthetic data generated via plug-in sampling. Staff developed this methodology for the cases when the original data follow either a multivariate normal or a multiple linear regression model. Staff established sufficient conditions under which our methodology will yield valid inference and studied properties of our methodology for the multiple linear regression model when certain conditions do not hold. Specifically, staff studied the scenario where the original data follow a linear regression model, and the data analyst observes a set of singly imputed synthetic data; however, the data generating model, imputation model, and data analysis model are not all the same. Our analysis includes both theoretical and empirical results to evaluate how statistical inference is affected. Staff also studied another scenario where the sufficient conditions for valid inference do not hold because the data producer uses the regression of y on x to generate synthetic data, but the data analyst's model is the regression of x on y. Under each of these scenarios, we compared the performance of our methodology for singly imputed synthetic data with the performance of established methods for multiply imputed synthetic data. Staff revised our manuscript, "Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models." All of the material discussed above is included in the revision. The manuscript was accepted for publication in the *Journal of Privacy and Confidentiality.*

Staff also completed the manuscript, "Inference for Multivariate Regression Model based on Synthetic Data generated under Fixed-Posterior Predictive Sampling: Comparison with Plug-in Sampling," which develops likelihood based inference based on singly and multiply imputed synthetic data, generated via fixed-posterior predictive sampling, under a multivariate linear regression model. The authors note that fixed-posterior predictive sampling differs from standard posterior predictive sampling. In this manuscript, they also provide some comparisons between fixed-posterior predictive sampling and plug-in sampling in terms of quality of inference and privacy protection. This manuscript was accepted for publication in *REVSTAT-Statistical Journal.*

Staff completed the manuscript, "Inference for Multivariate Regression Model based on Synthetic Data generated using Plug-in Sampling," which develops likelihood based inference based on both singly and multiply imputed synthetic data, generated via plug-in sampling, under a multivariate linear regression model. This manuscript was submitted for publication.

Staff began studying inference based on synthetic data when the data generating model, imputation model, and data analysis model are not all the same under posterior predictive sampling. Staff also began developing an approach to generate synthetic data using the conditional distribution of the data, given the sufficient statistics, under a multiple linear regression model.

Staff: Martin Klein (x37856), Bimal Sinha (CDAR), Thomas Mathew, Brett Moran

Summer at Census

Description: For each summer since 2009, recognized scholars in the following and related fields applicable to censuses and large-scale sample surveys are invited for short-term visits (one to five days) primarily between May and September: statistics, survey methodology, demography, economics, geography, social and behavioral sciences, and computer science. Scholars engage in collaborative research with Census Bureau researchers and staff and present a seminar based on their research.

Scholars are identified through an annual Census Bureau-wide solicitation by the Center for Statistical Research and Methodology.

Highlights: Staff facilitated all the details and background with staff from around the Census Bureau to host *2016 SUMMER AT CENSUS* with nearly forty scholars.

Staff: Tommy Wright (x31702), Michael Leibert

Research Support and Assistance

This staff provides substantive support in the conduct of research, research assistance, technical assistance, and secretarial support for the various research efforts.

Staff: Erica Magruder, Kelly Taylor

3. PUBLICATIONS

3.1 JOURNAL ARTICLES, PUBLICATIONS

Adragni, K.P., Al-Najjar, E., Martin, S., Popuri, S.K., and Raim, A.M. (2016). “Groupwise Sufficient Dimension Reduction with Principal Fitted Components,” *Computational Statistics.*

Abramowitz, J., O'Hara, B., and Morris, D.S. (In Press). “Risking Life and Limb: Estimating a Measure of Medical Care Economic Risk and Considering its Implications,” *Health Economics*.

Athreya, K.B. and Janicki, R. (2016). “Asymptotics of Powers of Binomial and Multinomial Probabilities,” *Statistics and Probability Letters, 112,* 58-62.

Blakely, C. and McElroy, T. (2016). “Signal Extraction Goodness-of-fit Diagnostic Tests Under Model Parameter Uncertainty,” *Econometrics Reviews,* 1-16.

Carden, S. and Livsey, J. (In Press). “Improved Policies Using Synthetic Data in Reinforcement Learning Algorithms,” *Intelligent Decision Technologies.*

Franco, C. and Bell, W. R. (2015). “Borrowing Information Over Time in Binomial/Logit Normal Models for Small Area Estimation,” Joint issue of *Statistics in Transition* and *Survey Methodology, 16 (4):* 563-584.

Holan, S., McElroy, T., and Wu, G. (In Press). “The Cepstral Model for Multivariate Time Series: The Vector Exponential Model,” *Statistica Sinica.*

Janicki, R. and McElroy, T. (2016). “Hermite Expansion and Estimation of Monotonic Transformations of Gaussian Data,” *Journal of Nonparametric Statistics, 28(1):* 207-234.

Klein, M. and Sinha, B. (2016). “Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models,” *Journal of Privacy and Confidentiality,7*: 43-98.

Lu, X. and West, D. (2016). “A New Proof that 4-connected Planar Graphs are Hamiltonian-connected,” *Discussiones Mathematicae Graph Theory, 36:* 555-564.

McElroy, T. (2016). “Multivariate Seasonal Adjustment, Economic Identities, and Seasonal Taxonomy,” *Journal of Business and Economics Statistics.* Published Online.

McElroy, T. (2016). “On the Measurement and Treatment of Extremes in Time Series,” *Extremes,* 1-24.

McElroy, T. (In Press). “Non-nested Model Comparisons for Time Series.” *Biometrika*.

McElroy, T. and Holan, S. (2016). “Estimation of Time Series with Multiple Long-Range Persistencies,” *Computational Statistics and Data Analysis, 101*: 44-56.

McElroy, T. and McCracken, M. (2016). “Multi-Step Ahead Forecasting of Vector Time Series,” *Econometrics Reviews,* 1-26.

McElroy, T. and Nagaraja, C. (2016). “Tail Index Estimation with a Fixed Tuning Parameter Fraction,” *Journal of Statistical Planning and Inference, 170:* 27-45.

Morris, D.S., Keller, A., and Clark, B. (2016). “An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census.” *Statistical Journal of the International Association for Official Statistics, 32(2)*: 177-188.

Morris, D.S., Schwarcz, D., and Teitelbaum, J.C. (In Press). “Do Credit-Based Insurance Scores Proxy for Income in Predicting Auto Claim Risk?” *Journal of Empirical Legal Studies.*

Moura, R., Klein, M., Coelho, C., and Sinha, B. (In Press). "Inference for Multivariate Regression Model based on Synthetic Data Generated under Fixed-Posterier Predictive Sampling: Comparison with Plug-in Sampling," *REVSTAT-Statistical Journal.*

Mulry, M. H., Nichols, E. M. and Childs Hunter, J. (2016). "A Case Study of Error in Survey Reports of Move Month Using the U.S. Postal Service Change of Address Records," *Survey Methods: Insights from the Field.* Retrieved from http://surveyinsights.org/?p=7794.

Mulry, M., Oliver, B., Kaputa, S., and Thompson, K. (In Press). "A Cautionary Note on Clark Winsorization," *Survey Methodology*.

Sellers, K.F., Morris, D.S., and Balakrishnan, N. (2016). "Bivariate Conway-Maxwell-Poisson Distribution: Formulation, Properties, and Inference," *Journal of Multivariate Analysis*, *150*: 152-168.

Sellers, K.F. and Raim, A.M. (2016). "A Flexible Zero-inflated Model to Address Data Dispersion," *Computational Statistics and Data Analysis*, *99*: 68-80.

Trimbur, T. and McElroy, T. (In Press). "Signal Extraction for Nonstationary Time Series with Diverse Sampling Rules," *Journal of Time Series Econometrics.*

Wildi, M. and McElroy, T. (2016). "Optimal Real-Time Filters for Linear Prediction Problems," *Journal of Time Series Econometrics, 8:*155-192.

Young, D.S., Raim, A.M., and Johnson, N.R. (In Press). "Zero-inflated Modelling for Characterizing Coverage Errors of Extracts from the U.S. Census Bureau's Master Address File," *Journal of the Royal Statistical Society: Series A.*

Zhu, L., Sellers, K.F., Morris, D.S., and Shmueli, G. (In Press). "Bridging the Gap: A Generalized Stochastic Process for Count Data", *The American Statistician,* http://dx.doi.org/10.1080/00031305.2016.1234976

3.2 BOOKS/BOOK CHAPTERS

Bell, W., Basel, W., and Maples, J. (2015). "An Overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) Program," in Monica Protesi (Ed.), *Analysis of Poverty Data by Small Area Methods,* London: Wiley.

Christen, P. and Winkler, W. E. (To Appear). "Record Linkage," in *Encyclopedia of Machine Learning and Data Mining*.

Erciulescu, A.L., Franco, C., and Lakini, P. (In Press). "Use of Administrative Records in Small Area Estimation," in A.Y. Chun and M. Larsen (Eds.), *Administrative Records for Survey Methodology,* Wiley Publishers.

Winkler, W. E. (2015). "Probabilistic Linkage," in Goldstein, H., Harron, K., and Dibbel, C. (Eds.), *Methodological Developments in Data Linkage*. Wiley.

3.3 PROCEEDINGS PAPERS

Joint Statistical Meetings, American Statistical Association, Seattle, Washington, August 8-13, 2015.
2015 Proceedings of the American Statistical Association

- Maria M. Garcia, Darcy Steeg Morris, and L. Kaili Diamond, "Implementation of Ratio Imputation and Sequential Regression Multivariate Imputation on Economic Census Products", 1056-1070.
- Martin Klein, Joanna Fane Lineback, and Joseph L. Schafer, "Evaluating Imputation and Estimation Procedures in a Survey of Wholesale Businesses", 1997-2008.
- Darcy Steeg Morris, Andrew Keller, and Brian Clark, "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census", 3278-3292.
- Mary Mulry and Andrew Keller, "Are Proxy Responses Better Than Administrative Records?" 2465-2479.
- Andrew M. Raim, Marissa N. Gargano, Nagaraj K. Neerchal, and Jorge G. Morel, "Bayesian Analysis of Overdispersed Binomial Data Using Mixture Link Regression", 2794-2808.

FCSM Proceedings, Federal Committee on Statistical Methodology Meeting, Washington, D.C., December 1-3, 2015.

- Laura Bechtel, Darcy Steeg Morris, and Katherine Jenny Thompson, "Using Classification Trees to Recommend Hot Deck Imputation Methods: A Case Study".

3.4 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY RESEARCH REPORTS

<http://www.census.gov/srd/csrmreports/byyear.html>

RR (Statistics #2015-04): Andrew Raim and Marissa N. Gargano. "Selection of Predictors to Model Coverage Errors in the Master Address File," December 30, 2015.

RR (Statistics #2016-01): Ryan Janicki. "Estimation of the Difference of Small Area Means from Different Time Periods," February 25, 2016.

RR (Statistics #2016-02): Osbert Pang and Brian C. Monsell. "Examining Diagnostics for Trading-Day Effects from X-13ARIMA-SEATS," March 11, 2016.

RR (Statistics #2016-03): Tommy Wright. "Two Optimal Exact Sample Allocation Algorithms: Sampling Variance Decomposition is Key," May 10, 2016.

RR (Statistics #2016-04): Mary H. Mulry and Andrew D. Keller, "Using 2010 Census Coverage Measurement Results to Compare Census Nonresponse Followup Proxy Responses with Administrative Records," August 30, 2016.

3.5 OTHER REPORTS

Hughes, T., Slud, E., Ashmead, R., and Walsh, R. (2016). "Results of a Field Pilot to Reduce Respondent Contact Burden in the American Community Survey's Computer Assisted Personal Interviewing Operation," *American Community Survey Research and Evaluation Report Memorandum Series, ACS16-RER-07.* http://www.census.gov/library/working-papers/2016/acs/2016_Hughes_01.html

Mulry, M., Clark, B., and Mule, T. (2016). "Final Report on the 2015 Census Test Evaluation Followup", *2020 Census Program Internal Memorandum Series: 2016.51.i.*

4. TALKS AND PRESENTATIONS

Statistics Colloquium, University of Maryland, Baltimore County, Baltimore, Maryland, October 2, 2015.

- Robert Ashmead, "Propensity Score Estimators for Causal Inference with Complex Survey Data."

2015 Morehouse Mathematics Fair, Morehouse College, Atlanta, Georgia, October 8, 2015.

- Kimberly Sellers, "Don't Count on Poisson: Introducing a Flexible Alternative Distribution to Model Count Data."

Colloquium Seminar, Biostatistics Department, Columbia University, New York City, New York, October 8, 2015.

- Tommy Wright, "Simple Exact Optimal Sample Allocation Algorithms, More Efficient Than Neyman Allocation: Sampling Variance Decomposition is Key."

Nielsen Office, Columbia, Maryland, *October 15, 2015.*

- Carolina Franco and William R. Bell, "Borrowing Information Over Time in Binomial/Logit Normal Models for Small Area Estimation."

Minisymposium Honoring Dianne O'Leary, SIAM Conference on Applied Linear Algebra, Atlanta, Georgia, October 26-30, 2015.

- Kimberly Sellers, "A Flexible Regression Model for Count Data."

Department of Statistics, University of Missouri, Columbia, Missouri, October 28, 2015.

- Martin Klein, "Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-normal Regression Samples."

NISS Workshop on Nonignorable Nonresponse, Bureau of Labor Statistics, Washington, D.C., November 12-13, 2015.

- Eric Slud, "Weighted Estimating Equations Based on Response Propensities in Terms of Covariates that are Observed only for Responders."

Workshop on Data Integration and Applications at the IEEE International Conference on Data Mining, Atlantic City, New Jersey, November 14, 2015.

- William E. Winkler, "Keynote: Clean-up and Preliminary Analysis for Data Mining Sets of National Files."

Time Series Workshop, Office of National Statistics, London, United Kingdom, November 19-20, 2015.

- Brian Monsell and James Livsey, "Overview of Time Series Issues and Research at the U.S. Census Bureau."
- Brian Monsell, "Weekly Seasonal Adjustment."
- James Livsey, "Diagnostics for Deciding on Moving Holiday Window."

Federal Committee on Statistical Methodology Research Conference, Washington, D.C., December 1, 2015.

- Carolina Franco, Roderick J. Little, Thomas A. Louis, and Eric V. Slud, "Comparative Study of Confidence Intervals for Proportions in Complex Surveys."

Departmental Seminar, Department of Statistics, University of Kentucky, Lexington, Kentucky, December 4, 2015.

- Kimberly Sellers, "A Flexible Regression Model for Count Data."

Computational and Financial Econometrics, London, United Kingdom, December 12-14, 2015.

- Tucker McElroy, "Seasonal Adjustment of Meager Time Series."

8th International Conference of the ERCIM Working Group on Computational and Methodological Statistics (CMStatistics 2015), University of London, London, United Kingdom. December 13, 2015.

- Carolina Franco, Serena Arima, William R. Bell, Gauri Datta, and Brunero Liseo, "Bayesian Treatment of a Multivariate Fay-Herriot Functional Measurement Error Model with Applications."

Universidad Carlos III de Madrid, Madrid, Spain, December 19, 2015

- Carolina Franco and William R. Bell, "Temporal Extensions to a Hierarchical Model for Proportions from Complex Survey Data. Statistical Seminar."

Joint Program in Survey Methodology, University of Maryland, February 5, 2016

- William E. Winkler, "Clean-up and Preliminary Analysis of Sets of National Files."

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, Maryland, February 19, 2016.

- Andrew Raim, "An Extension of Generalized Linear Models to Finite Mixture Outcomes."

Department of Statistics Colloquium, University of Kentucky, Lexington, Kentucky, February 19, 2016.

- Tommy Wright, "The Equivalence of Neyman Optimum Allocation for Sampling and Equal Proportions for Apportioning the U.S. House of Representatives."

2016 Ross-Royall Symposium. Johns Hopkins University, Baltimore, Maryland. February 26, 2016.

- William R. Bell and Carolina Franco, "Combining Estimates from Related Surveys via Bivariate Models."

Statistics Colloquium, University of Maryland, College Park, Maryland, March 4, 2016.

- Emanuel Ben-David, "Gaussian DAG models with Symmetries"

24th Symposium on Nonlinear Dynamics and Econometrics, University of Alabama, Tuscaloosa, Alabama, March 11, 2016.

- Thomas Trimbur (with Bill Bell), "The Effects of Seasonal Heterskedasticity in Time Series on Trend Estimation and Seasonal Adjustment."

Cameroon International Conference on Recent Developments in Applied Statistics, Yaounde, Cameroon, March 14-18, 2016.

- Tommy Wright (Keynote Address), "No Calculations When Observations Can Be Made."

17th Annual OxMetrics User Conference, The George Washington University, Washington, D.C., March 18, 2016.

- Thomas Trimbur (with Bill Bell), "The Effects of Seasonal Heterskedasticity in Time Series on Trend Estimation and Seasonal Adjustment."

Statistics Canada Methodology Symposium, Gatineau, Quebec, Canada, March 22 – 24, 2016.

- Mary Mulry, Elizabeth M. Nichols, and Jennifer Hunter Childs, "Using Administrative Records to Evaluate Survey Data."

Business Week, University of Texas at Arlington, Arlington, Texas, March 28 - April 1, 2016.

- Mary Mulry. "Statistical Methods Used in Planning and Implementing the 2010 Census Communications Campaign."

Work Outside the Book: Humanities Career Tracks Outside of Academia, University of Maryland, College Park, Maryland, April 6, 2016.

- Lauren Emanuel, "English Majors in the Federal Government."

SAMSI Workshop on Games and Decisions in Reliability and Risk, Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina, May 18, 2016.

- Kimberly Sellers, "A Generalized Statistical Control Chart for Over- or Under-dispersed Data."

10th Annual Probability & Statistics Day at UMBC, Baltimore County, Maryland, May 20-21, 2016.

- Martin Klein, "Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multiple Linear Regression Model."
- Tommy Wright, "Measurement for Official Statistics."

Department of Biostatistics Seminar, University of Iowa, Iowa City, Iowa, June 17, 2016.

- Tommy Wright, "The Equivalence of Neyman Optimum Allocation for Sampling and Equal Proportions for Apportioning the U.S. House of Representatives."

Iowa Summer Institute in Biostatistics, University of Iowa, Iowa City, Iowa, June 17, 2016.

- Tommy Wright, "Measurement and Official Statistics."

Fifth International Conference on Establishment Surveys, Geneva, Switzerland, June 21-24, 2016.

- Mary Mulry, Stephen Kaputa, and Katherine J. Thompson, "Setting Parameters for M-estimation."

Summer Program in Research and Learning (SPIRAL), Morgan State University, Baltimore, MD, July 18, 2016

- Kimberly Sellers, “Don’t Count on Poisson: Introducing the Conway-Maxwell-Poisson Distribution.”

Joint Statistical Meetings, American Statistical Association, Chicago, Illinois, July 31-August 4, 2016.

- Robert Ashmead and Eric Slud, “Inference from Complex Survey-Embedded Field Experiments.”
- Gauri Datta, “A Bayesian Generalized CAR Model for Correlated Signal Detection.”
- Aaron Gilary, Yang Cheng, and Eric Slud, “An Overview of Current Population Survey Variance Methodology.”
- Christopher Hassett, Scott Holan, and Tucker McElroy, “A Bayesian Approach to Multivariate Signal Extraction.”
- Krista Heim and Andrew Raim, “Predicting Coverage Error on the Master Address File Using Spatial Modeling Methods at the Block Level.”
- Patrick Joyce, “Evaluation of Estimation Methods for Section 203 of the Voting Rights Act.”
- Joanna Lineback, Martin Klein, and Joseph Schafer, “Exploring New Estimation Techniques for the Monthly Wholesale Trade Survey.”
- James Livsey, Tucker McElroy, and Anindya Roy, “Residual Diagnostics for Automatic Model Selection.”
- Bo Lu and Robert Ashmead, “Causal Inference with Unequal Sampling Weights: Investigating Policy Effect Using Population Health Surveys.”
- Jerry Maples, “Estimating Design Effects in Small Areas and Domains by Aggregation of Domains/Areas.”
- Brian Monsell, “An Examination of Weekly Seasonal Adjustment.”
- Vincent Mule, Andrew Keller, and Darcy Morris, “Using Administrative Records to Identify Occupied and Vacant Units.”
- Mary Mulry, Tom Mule, and Brian Clark, “Using the 2015 Census Test Evaluation Follow-Up to Compare Nonresponse Follow-Up with Administrative Records.”
- Senthilkuman Muthiah, Eric Slud, Mihai Pop, and Hector Bravo, “Dimensional Reduction of Metogenomic Data with Ecological Equivalence.”
- Osbert Pang, Brian Monsell, William Bell, and James Livsey, “Accommodating Weather Effects in Seasonal Adjustment.”
- Brandon Park, Anand Vidyashankar, Tucker McElroy, and Jie Xu, “Supervised Implicit Network Construction and Analysis of Related Network-Wide Metrics.”
- Jie Peng, Kalimuthu Krishnamoorthy, and Thomas Mathew, “A Simple Method for Assessing Occupational Exposure via the One-Way Random Effects Model.”
- Andrew Raim, “Informing Maintenance to the U.S. Census Bureau’s Master Address File with Statistical Decision Theory.”
- Anindya Roy and Tucker McElroy, “Test Based on Frobenius Norm Distance of Spectral Matrices for Presence of Structural Components.”
- Kimberly Sellers, Darcy Morris, and Narayanaswamy Balakrishnan, “Introducing the Bivariate Conway-Maxwell-Poisson Distribution.”
- Eric Slud and Robert Ashmead, “Design of Sample Surveys That Complement Observational Data to Achieve Population Coverage.”
- Yves Thibaudeau and Darcy Morris, “Bayesian Decision Theory for Further Optimizing the Use of Administrative Records in the Census NRFU.”
- Thomas Trimbur and William Bell, “The Effects of Seasonal Heteroskedasticity on Trend Estimation and Seasonal Adjustment for Time Series.”
- William Winkler, “Quality and Analysis of Sets of Files.”

ODRS 2016: Conference on Ordered Data and their Applications in Reliability and Survival Analysis, Session: Discrete Distributions, McMaster University, Hamilton, Ontario, Canada, August 8, 2016.

- Kimberly Sellers, “Bivariate Conway-Maxwell-Poisson Distribution: Formulation, Properties, and Inference.”

Small Area Estimation Conference, Maastricht University, Maastricht, Netherlands, August 18, 2016.

- Carolina Franco, Discussant for Keynote Address by Jiming Jiang, “Classified Mixed Model Prediction and Small Area Estimation.”

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

William Winkler, U.S. Census Bureau, “Edit/Imputation Course,” October 1, 2015.

William Winkler and Edward Porter, U.S. Census Bureau, “Record Linkage Course,” October 20 & 21, 2015.

William Winkler, U.S. Census Bureau, “Quality and Analysis or Sets of National Files,” November 3, 2015.

Jared Murray, Carnegie Mellon University, “Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence,” November 19, 2015.

Bret Hanlon, University of Wisconsin-Madison, “Robust Estimation for a Supercritical Branching Processor under Family-Size Sampling,” December 8, 2015.

Jason Bernstein, The Pennsylvania State University, “Time Series Analysis of Motor Proteins,” January 12, 2016.

Zachary Seeskin, Northwestern University, “Effects of Census Accuracy on Apportionment of Congress and Allocations of Federal Funds,” January 21, 2016.

Steve Carden, Georgia Southern University, “Reinforcement Learning and Marginalized Transition Models,” March 24, 2016.

Emanuel Ben-David, U.S. Census Bureau, “An Introduction to Probabilistic Graphical Models,” March 29, 2016.

Yves Thibaudeau and William Winkler, U.S. Census Bureau, “Edit & Imputation Theory & Computational Algorithms,” April 6, 13, & 20, 2016.

Douglas Galagate, U.S. Census Bureau/University of Maryland, College Park, “Causal Inference with a Continuous Treatment and Outcome: Alternative Estimators for Parametric Dose-Response Functions,” May 10, 2016.

Yinglei Lai, The George Washington University, *SUMMER AT CENSUS*, “Exploration of Concordant Changes among Multiple Data Sets,” May 17, 2016.

Subhashis Ghoshal, North Carolina State University, *SUMMER AT CENSUS*, “Bayesian Nonparametric Methods for Data-Science,” May 18, 2016.

Randall Akee, University of California, Los Angeles, *SUMMER AT CENSUS*, “Land Titles and Dispossession: Allotment on American Indian Reservations,” May 24, 2016.

Philip L.H. Yu, The University of Hong Kong, *SUMMER AT CENSUS*, “Rank Aggregation Using Distance-Based Models,” May 25, 2016.

Domingo Morales, University of Miguel Hernández of Elche, Spain, *SUMMER AT CENSUS*, “Multivariate Fay-Herriot Models for Small Area Estimation,” May 31, 2016.

Domingo Morales, University of Miguel Hernández of Elche, Spain, *SUMMER AT CENSUS*, “Small Area Estimation of Non-Linear Parameters Under a Two-Fold Nested Error Regression Model,” June 1, 2016.

Jae-Kwang Kim (ASA/NSF/Census Research Fellow), Iowa State University, “Some Recent Topics on Informative Sampling,” June 2, 2016.

James D. Wilson, University of San Francisco, “A Significance-based Community Extraction Method for Multilayer Networks,” June 2, 2016.

Zachary Seeskin, (U.S. Census Bureau Dissertation Fellow), Northwestern University, “Evaluating the Use of Commercial Data to Improve Survey Estimates of Property Taxes,” June 7, 2016.

John Iceland, The Pennsylvania State University, *SUMMER AT CENSUS*, “Did We Win the War on Poverty? No, but…,” June 7, 2016.

Wendy Manning, Bowling Green State University, *SUMMER AT CENSUS*, “Measuring Cohabitation in National Surveys,” June 8, 2016.

Sheela Kennedy, University of Michigan, *SUMMER AT CENSUS*, “The Changing Transition to Adulthood in the U.S.: Trends in Demographic Role Transitions and Age Norms since 2000,” June 9, 2016.

Sharon Sassler, Cornell University, *SUMMER AT CENSUS*, “A Cross-National Comparison of the Consequences of Partnered Childbearing for Mother’s Mid-Life Health,” June 9, 2016.

Vitaly Shmatikov, Cornell Tech, *SUMMER AT CENSUS*, “Machine Learning and Privacy: Friends or Foes?” June 14, 2016.

Henry Schneider, Cornell University, *SUMMER AT CENSUS*, “Promoting Best Practices in a Multitask Workplace: Experimental Evidence on Checklists,” June 14, 2016.

Carolyn Liebler, University of Minnesota, *SUMMER AT CENSUS*, “The Occupational Structure of the American Indian and Alaska Native Workforce,” June 20, 2016.

Christoph Sax, Christoph Sax Data Analytics LLC, *SUMMER AT CENSUS*, “R-Development: User Interfaces and Package Creation,” June 21, 2016.

Zhuoqiong He, University of Missouri-Columbia, *SUMMER AT CENSUS*, “Assessing and Adjusting Nonresponse Bias in Small Area Estimation via Bayesian Hierarchical Spatial Models,” June 21, 2016.

Elizabeth Fussell, Brown University, *SUMMER AT CENSUS*, “Disasters and Residential Change in the U.S., 1997-2013: Migrants’ Reasons for Moving, Socio-Demographic Selection, and Residential Outcomes,” June 27, 2016.

Ashley Amaya (U.S. Census Bureau Dissertation Fellow, University of Maryland), RTI International, “Enhancing the Understanding of the Relationship between Social Integration and Nonresponse,” June 28, 2016.

Narayan Sastry, University of Michigan, *SUMMER AT CENSUS*, “The Effects of Hurricane Katrina on the New Orleans Population: Results from the American Community Survey,” June 28, 2016.

Scott Holan, University of Missouri, *SUMMER AT CENSUS*, “Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics,” June 29, 2016.

Bikas K. Sinha, Indian Statistical Institute, Kolkata, *SUMMER AT CENSUS*, “Randomized Response & A New Hartely-Politz-Simmons Technique,” July 12, 2016.

Don Dillman, Washington State University, *SUMMER AT CENSUS*, “The Promises and Perils of Web-Push Methodologies,” July 12, 2016.

Deirdre Giesen, Statistics Netherlands, *SUMMER AT CENSUS*, “Response Burden in Official Business Surveys: Relevance, Concepts & Measurement,” July 13, 2016.

Deirdre Giesen, Statistics Netherlands, *SUMMER AT CENSUS*, “Management of Response Burden in Official Surveys,” July 14, 2016.

Joe Murphy, Research Triangle Institute (RTI) International, *SUMMER AT CENSUS*, “Methods and Technology for Monitoring Survey Data Quality during Data Collection,” July 14, 2016.

Thurston Domina, University of North Carolina, *SUMMER AT CENSUS*, “Beyond Tracking and Detracking: The Dimensions of Organizational Differentiation in Schools,” July 18, 2016.

Bernard Black, Northwestern University, *SUMMER AT CENSUS*, “The Effect of Health Insurance on Near-Elderly Health and Mortality,” July 19, 2016.

Emily Penner, University of California, Irvine, *SUMMER AT CENSUS*, “The Causal Effects of Cultural Relevance: Evidence from an Ethnic Studies Curriculum,” July 19, 2016.

Donald Rubin, Harvard University, *SUMMER AT CENSUS*, “A New Class of Models for Missing Data,” July 20, 2016.

Susie Fortier, Statistics Canada, *SUMMER AT CENSUS*, “Various Algorithmic Approaches for the Balancing Problem,” July 21, 2016.

David Haziza, University of Montreal, *SUMMER AT CENSUS*, “Multiply Robust Imputation Procedures for the Treatment of Item Nonresponse in Surveys,” July 21, 2016.

Malte Schierholz, German Institute for Employment Research, “New Methods for the Measurement of Occupation,” July 26, 2016.

Wolfgang Keller, University of Colorado, *SUMMER AT CENSUS*, “International Trade and Job Polarization: Evidence at the Worker Level,” July 26, 2016.

Alexander Bartik, Massachusetts Institute of Technology, *SUMMER AT CENSUS*, “Winners and Losers from Productivity and Amenity Changes: Evidence from Longitudinal Census Data and a Natural Resource Boom,” July 27, 2016.

Marti Hearst, University of California, Berkeley, *SUMMER AT CENSUS*, “Seeking Simplicity in Search User Interface,” August 2, 2016.

Andre Kurman, Drexel University, *SUMMER AT CENSUS*, “Downward Wage Rigidity in the United States,” August 2, 2016.

Marti Hearst, University of California, Berkeley, *SUMMER AT CENSUS*, “DesignITRight: How and Why to Integrate User-Centered Design into All Phases of IT Development,” August 3, 2016.

Daniel Goldberg, Texas A&M University, *SUMMER AT CENSUS*, “Geocomputational Approaches for Geocoding and Addressing,” August 8, 2016.

Werner Kuhn, University of California-Santa Barbara, *SUMMER AT CENSUS*, “Exploring the Notion of Spatial Data Lenses,” August 9, 2016.

Anand Vidyashanker, George Mason University, *SUMMER AT CENSUS*, “Implicit Networks in High Dimensional Problems,” August 10, 2016.

Roberto Rigobon, Massachusetts Institute of Technology (MIT), *SUMMER AT CENSUS*, “Big Data in Economic Measurement,” August 22, 2016.

Fang Qiu, University of Texas at Dallas, *SUMMER AT CENSUS*, “Curve Matching Approaches to Waveform Classification: A Case Study Using ICESat,” September 13, 2016.

Jennifer Van Hook, The Pennsylvania State University, *SUMMER AT CENSUS*, “Moving to the Land of Milk and Cookies: How Migration and Settlement in the U.S. Shapes Children’s Diets,” September 20, 2016.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

6.2 SIGNIFICANT SERVICE TO PROFESSION

Robert Ashmead

- Refereed papers for *Journal of Survey Statistics and Methodology*, *Journal of Official Statistics,* and *The American Statistician*

Emanuel Ben-David

- Refereed papers for the 33rd International Conference on Machine Learning (ICML 2016), 19th International Conference on Artificial Intelligence and Statistics (AISTAT 2016), *Annals of Applied Statistics*, *Mathematical Reviews*, *Journal of Statistical Planning and Inference*, *The American Statistician*, *Statistica Sinica,* and *Journal of Statistical Planning*

Carolina Franco

- Refereed papers for *The American Statistician* and the *Journal of Official Statistics*

Maria Garcia

- Member, Program Committee, Federal Committee on Statistical Methodology (FCSM) Research Conference, December 2015
- Organizer and Session Chair, FCSM Session: "Methods for Missing Data Imputation"
- Organizer, FCSM Session: "Imputation, Multiple Imputation, and Administrative Records"

Patrick Joyce

- Refereed a paper for *Statistical Science*

Jerry Maples

- Refereed a paper for *Journal of Official Statistics*

Martin Klein

- Refereed papers for *Journal of the International Association for Official Statistics, Journal of the Royal Statistical Society-Series A*, and *Statistical Papers*
- Member, Ph.D. Dissertation in Statistics Committee, University of Maryland, Baltimore County
- Session Chair, 10th Annual Probability & Statistics Day, University of Maryland, Baltimore County

James Livsey

- Organizer, 2016 JSM Session: “Time Series Seasonal Adjustment: Weekly Valued and Weather Adjustments”
- Organizer, 2016 JSM Session: “Time Series Modeling: Seasonality, Multivariate, and Testing”
- Chair, 2016 JSM Session: “Recent Advances and Applications of Spatial and Spatio-Temporal Models for Official Statistics”
- Refereed papers for *Computational Statistics, Journal of Time Series Econometrics, Environmetrics, Applied Stochastic Models in Business and Industry,* and *Sankhya, Series B.*

Thomas Mathew

- Associate Editor, *Journal of the American Statistical Association*
- Associate Editor, *Statistical Methodology*
- Associate Editor*, Sankhya, Series B*
- Editorial Board, Member, *Journal of Occupational and Environmental Hygiene*
- Member, American Statistical Association’s Committee on W.J. Youden Award in Inter-laboratory Testing
- Refereed articles for *Journal of the American Statistical Association, Journal of Official Statistics, Sankhya, Statistics and Probability Letters* and *Communications in Statistics*

Tucker McElroy

- Refereed papers for *Annals of Statistics, Journal of Applied Econometrics, Journal of Official Statistics,* and *Communications in Statistics*

Brian Monsell

- Chair, 2016 JSM Session: “Forecasting and ARMA Modeling”

Darcy Morris

- Chair, 2016 JSM Session: “Modeling Multivariate Count Data: Multivariate Extensions and Generalizations of Standard Count Distributions”
- Refereed a paper for *CityScape*

Mary Mulry

- Associate Editor, *Journal of Official Statistics*
- Methodology co-Editor, *Statistical Journal of the International Association of Official Statistics*
- Refereed a paper for *Journal of Survey Statistics and Methodology*
- Invited Session Organizer, Fifth International Conference on Establishment Surveys (ICES-V)
- Chair, 2016 JSM Session: “Tackling the Challenges of Missing Data in Surveys: Applying Methods and Assessing Uncertainty”

Osbert Pang

- Refereed a paper for *Biometrika*

Ned Porter

- Reviewed papers for 22nd Association of Computing Machinery, Conference for Knowledge Discovery and Data Mining Applied Data Science Tract

Andrew Raim

- Chair, 2016 JSM Session: “Modeling Multivariate Court Data: Multivariate Extensions and Generalizations of Standard Count Distributions”
- Refereed papers for *The American Statistician* and *Statistics and Operations Research Transactions*
- Member, Ph.D. Dissertation in Statistics Committee, University of Maryland, Baltimore County
- Presented a Workshop on High Performance R and Big Data at the University KMUTT in Bangkok, Thailand with Nagaraj Neerchal (U. of Maryland, Baltimore County) and George Ostrouchov (Oak Ridge National Laboratory)

Kimberly Sellers

- Member, American Statistical Association Committee on Women in Statistics
- Associate Editor, *The American Statistician*
- Advisory Board Member and Director, BDN STEMers for International Black Doctoral Network Association, Incorporated
- Refereed papers for *Applied Stochastic Models in Business and Industry, Biometrics, Communications in Statistics – Theory and Methods, Computers & Industrial Engineering, Lifetime Data Analysis, Quality and Reliability Engineering International,* and *Statistics*
- Member, Scientific Program Committee, International Conference on Statistical Distributions and Applications (ICOSDA) 2016
- Organizer, 2016 ICOSDA Invited Session: “Don’t Count on Poisson! Introducing the Conway-Maxwell-Poisson Distribution for Statistical Methodology Regarding Count Data”

Eric Slud

- Associate Editor, *Biometrika*
- Associate Editor, *Journal of Survey Statistics and Methodology*
- Associate Editor, *Lifetime of Data Analysis*
- Chair, 2016 JSM Session: “Nonresponse Adjustment and Nonresponse Bias Reduction Methods”
- Discussant, 2016 JSM Session: “Resampling Methods in Mixed Effects Models with Applications in Small Area Estimation and Other Related Fields”

Yves Thibaudeau

- Refereed papers for *Statistics in Medicine* and *The Journal of Official Statistics*

William Winkler

- Refereed papers for the *Journal of Official Statistics, Journal of Survey Statistics and Methodology*, *JASA* and *Statistical Data Protection 2016*
- Reviewer, grant proposal and made recommendation related to large grant proposal by the Dutch government
- Reviewer, grant proposal on record linkage for the National Science Foundation
- Associate Editor, *Journal of Privacy and Confidentiality*
- Associate Editor, *Transactions on Data Privacy*
- Member, Program Committee for *Statistical Data Protection 2016*
- Member, Program Committee for *IEEE 2015 ICDM Data Integration and Applications* and *IEEE 2016 ICDM Data Integration and Applications*
- Member, Program Committee for ACM Workshop on Population Informatics at KDD'16
- Member, Statistics Ph.D. Committee at the University of Maryland

Tommy Wright

- Associate Editor, *The American Statistician*
- Chair, Waksberg Award Committee, *Survey Methodology*
- Member, Board of Trustees, National Institute of Statistical Sciences
- Reviewer, Tenure Review of Faculty Member, Biostatistics Department, Columbia University

6.3 PERSONNEL NOTES

Alisha Armas completed graduate studies at American University and accepted another position.

Dan Weinberg (new Ph.D., Mathematics, University of Maryland, College Park) joined our Time Series Research Group.

Rolando Rodriguez accepted a position in the Center for Disclosure Avoidance.

Bret Hanlon joined our Simulation, Modeling, and Data Visualization Research Group.

Adam Maidman (Ph.D student in statistics at University of Minnesota) joined our center as a summer intern.

Jae-Kwang Kim (Statistics Professor at Iowa State University) joined the Census Bureau as an ASA/NSF/Census Research Fellow.

Claire Bowen (Ph.D. student in statistics at Notre Dame) joined our center as an NSF Graduate Research Intern.

APPENDIX A **Center for Statistical Research and Methodology FY 2016**
Program Sponsored Projects/Subprojects With Substantial Activity and Progress and Sponsor Feedback
(Basis for PERFORMANCE MEASURES)

Project #	Project/Subproject Sponsor(s)	CSRM Contact	Sponsor Contact
	DECENNIAL		
6650B23	Redesigning Field Operations		
6750B01	Administrative Records Data		
6550B01	Data Coding, Editing, and Imputation		
6250B07	Policy		
	1. Decennial Record Linkage	William Winkler	Tom Mule
	2. Coverage Measurement Research	Jerry Maples	Tim Kennel
	3. Analysis of the 2015 Census Test Evaluation Follow-up	Mary Mulry	Tom Mule
	4. Record Linkage Error-Rate Estimation Methods	William Winkler	Tom Mule
	5. Supplementing and Supporting Non-Response with Administrative Records	Michael Ikeda	Tom Mule
	6. Identifying "Good" Administrative Records for 2020 Census NRFU Curtailment Targeting	Darcy Steeg Morris	Tom Mule
	7. Special Census: Disclosure Avoidance in Group Quarters	Rolando Rodriguez	Michael Freiman
6350B02	Address Canvassing In Field		
	8. Master Address File (MAF) Error Model and Quality Assessment	Andrew Raim	Laura Ferreira
	9. Development of Block Data Tracking Database	Tom Petkunas	Michael Ratcliffe
6385B70	American Community Survey (ACS)		
	10. ACS Applications for Time Series Methods	Tucker McElroy	Mark Asiala
	11. Data Analysis of ACS CATI-CAPI Contact History	Eric Slud	Elizabeth Poehler
	12. Confidence Intervals for Proportions in ACS Data	Carolina Franco	Mark Asiala
	13. Voting Rights Section in 203 Model Evaluation and Enhancements Towards Mid-Decadal Determinations	Patrick Joyce	James Whitehorne
	DEMOGRAPHIC		
0906/1444X00	Demographic Surveys Division (DSD) Special Projects		
	14. Data Integration	Ned Porter	Christopher Boniface
TBA	Population Division Special Projects		
	15. Introductory Sampling Workshop	Tommy Wright	Oliver Fischer
7165016	Social, Economic, and Housing Statistics Division Small Area Estimation Projects		
	16. Research for Small Area Income and Poverty Estimates (SAIPE)	Jerry Maples	Wes Basel
	17. Small Area Health Insurance Estimates (SAHIE)	Ryan Janicki	Wes Basel
	18. Sub-County Estimates of Poverty from Multi-year ACS Data	Jerry Maples	Wes Basel
	ECONOMIC		
1183X01	Economic Statistical Collection		
	19. Research on Imputation Methodology for the Monthly Wholesale Trade Survey	Martin Klein	Joe Schafer
	20. Use of Big Data for Retail Sales	Darcy Steeg Morris	Rebecca Hutchinson
2220B10	Economic Census/Survey Engineering: Time Series Research; Economic Missing Data/Product Line Data; Development/SAS		
	21. Seasonal Adjustment Support	Brian Monsell	Kathleen McDonald-Johnson
	22. Seasonal Adjustment Software Development and Evaluation	Brian Monsell	Kathleen McDonald-Johnson
	23. Research on Seasonal Time Series: Modeling & Adjustment Issues	Tucker McElroy	Kathleen McDonald-Johnson
	24. Supporting Documentation & Software: X-12-ARIMA & X-13A-S	Brian Monsell	Kathleen McDonald-Johnson
	25. Missing Data Adjustment Methods for Product Data in the Economic Census	Darcy Steeg Morris	Jenny Thompson
7103012	2012 Commodity Flow Survey		
	26. 2012 Commodity Flow Survey	Robert Ashmead	Joanna Fane Lineback
TBA	Investigation of Alternative Methods for Resolving Balance Complex Failures in StEPS		
	27. Investigation of Alternative Methods for Resolving Balance Complex Failures in StEPS	Maria Garcia	Laura Bechtel
	RESEARCH AND METHODOLOGY DIRECTORATE		
TBA	*28. Business Dynamics Statistics—Export File Weighting Issue*	Maria Garcia	Fariha Kamal
	ADMINISTRATION AND CFO PROJECT		
TBA	*29. Assessment of Census Bureau's Finance Methodology for Estimating Accruals*	Tommy Wright	Robin Guinn
	CENSUS BUREAU		
7236045	*30. National Survey of Drug Use & Health*	Robert Ashmead	J.D. Wynn

APPENDIX B

FY 2016 PROJECT PERFORMANCE MEASUREMENT QUESTIONNAIRE

CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY

Dear

In a continuing effort to obtain and document feedback from program area sponsors of our projects or subprojects, the Center for Statistical Research and Methodology will attempt for the eighteenth year to provide *seven measures of performance*, this time for the fiscal year 2016. For FY 2016, the *measures of performance* for our center are:

Measure 1. Overall, Work Met Expectations: Percent of FY 2016 Program Sponsored Projects/Subprojects where sponsors reported that work met their expectations.
Measure 2. Established Major Deadlines Met: Percent of FY 2016 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met.
Measure 3a. At Least One Improved Method, Developed Technique, Solution, or New Insight: Percent of FY 2016 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight.
Measure 3b. Plans for Implementation: Of the FY 2016 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight, the percent with plans for implementation.
Measure 4. Predict Cost Efficiencies: Number of FY 2016 Program Sponsored Projects/Subprojects reporting at least one "predicted cost efficiency."
Measure 5. Journal Articles, Publications: Number of journal articles (peer review) and publications documenting research that appeared or were accepted in FY 2016.
Measure 6. Proceedings Publications: Number of proceedings publications documenting research that appeared in FY 2016.

These measures will be based on response to the five questions on this form from our sponsors as well as from members of our center and will be used to help improve our efforts.

To construct these seven measures for our center, we will combine the information for all of our program area sponsored projects or subprojects obtained during December 1 thru December 9, 2016 using this questionnaire. Your feedback is requested for:

Project Number and Name: ______________
Sponsoring Division(s): ________________

After all information has been provided, the CSRM Contact ____________ will ensure that the signatures are obtained in the order indicated on the last page of this questionnaire. We very much appreciate your assistance in this undertaking.

____

Tommy Wright Date
Chief, Center for Statistical Research and Methodology

Brief Project Description ***(CSRM Contact will provide from Division's Quarterly Report):***

Brief Description of Results/Products from FY 2016 ***(CSRM Contact will provide)****:*

(over)

TIMELINESS:
Established Major Deadlines/Schedules Met

1(a). Were all established major deadlines associated with this project or subproject met? **(Sponsor Contact)**

□ Yes □ No □ No Established Major Deadlines

1(b). If the response to 1(a) is No, please suggest how future schedules can be better maintained for this project or subproject. **(Sponsor Contact)**

QUALITY & PRODUCTIVITY/RELEVANCY:
Improved Methods / Developed Techniques / Solutions / New Insights

2. Listed below are at most 2 of the top improved methods, developed techniques, solutions, or new insights offered or applied on this project or subproject in FY 2016 where an CSRM staff member was a significant contributor. Review “a” and “b” below **(provided by CSRM Contact)** and make any additions or deletions as necessary. For each, please indicate whether or not there are plans for implementation. If there are no plans for implementation, please comment.

□ No improved methods/techniques/solutions/new insights developed or applied.

□ Yes as listed below. (See a and b.)

Plans for Implementation?

a. ____________________________ Yes □ No □
____
____
____
____

b. ____________________________ Yes □ No □
____
____
____
____

Comments (Sponsor Contact):

COST:
Predict Cost Efficiencies

3. Listed **(provided by CSRM Contact)** below are at most two research results or products produced for this project or subproject in FY 2016 that predict cost efficiencies. Review the list, and make any additions or deletions as necessary. Add any comments.

□ No cost efficiencies predicted.
□ Yes as listed below. (See a and b.)

a.

b.

Comments (Sponsor Contact):

OVERALL:
Expectations Met/Improving Future Communications

4. Overall, work on this project or subproject by CSRM staff during FY 2016 met expectations. **(Sponsor Contact)**

□ Strongly Agree
□ Agree
□ Disagree
□ Strongly Disagree

5. Please provide suggestions for future improved communications or any area needing attention on this project or subproject. **(Sponsor Contact)**

(CSRM Contact will coordinate the signatures as noted and pass to CSRM Chief.)

First____
Sponsor Contact Signature Date

Second____
CSRM Contact Signature Date

Center for Statistical Research and Methodology

Research & Methodology Directorate

STATISTICAL COMPUTING AREA

Bill Winkler (Acting)
VACANT

Machine Learning & Computational Statistics Research

Bill Winkler
Emanuel Ben-David
Xiaoyun Lu

Missing Data Methods Research

Yves Thibaudeau
Douglas Galagate (S)
Maria Garcia
Darcy Morris
Jun Shao (U. of WI)

Research Computing Systems & Applications

Chad Russell
Tom Petkunas
Ned Porter

Simulation, Modeling, & Data Visualization Research

Martin Klein
Claire Bowen (NSF-GRIP)
Isaac Dompreh
Brett Hanlon
Brett Moran
Nathan Yau (FLOWINGDATA.COM)

MATHEMATICAL STATISTICS AREA

Eric Slud
Erica Magruder (HRD)

Sampling & Estimation Research

Eric Slud (Acting)
Robert Ashmead
Mike Ikeda
Patrick Joyce
Mary Mulry

Small Area Estimation Research

Jerry Maples
Gauri Datta (U. of GA)
Carolina Franco
Ryan Janicki

Time Series Research

Brian Monsell
Osbert Pang
Tucker McElroy
James Livsey
Aninyda Roy (UMBC)
Thomas Trimbur
Dan Weinberg

Experimentation & Modeling Research

Tommy Wright (Acting)
Thomas Mathew (UMBC)
Andrew Raim
Kimberly Sellers (Georgetown U.)

Tommy Wright, Chief
Kelly Taylor
Lauren Emanuel
Jae-Kwang Kim (F)
Michael Hawkins
Michael Leibert
Andrew Perry (S)

(S) Student
(F) ASA/NSF/Census Research Fellow
(NSF-GRIP) NSF Graduate Research Internship Program

September 30, 2016