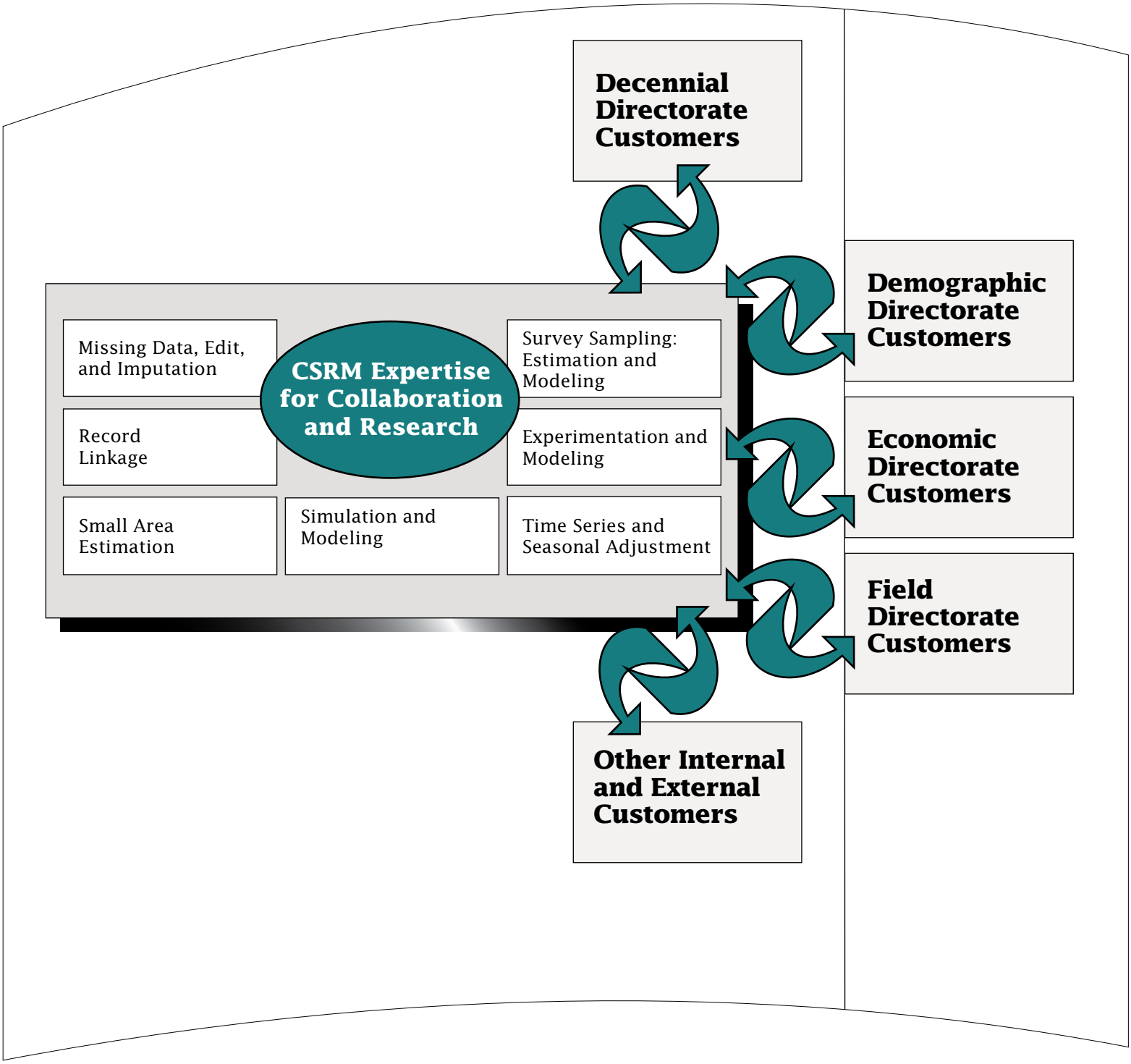
Annual Report *of the* Center for Statistical Research and Methodology

Research and Methodology Directorate

Fiscal Year 2015



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

United States™ Census Bureau

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, nonresponse, etc.) 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 2015, this report is an accounting of our work and our results.

Center for Statistical Research & Methodology

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 2015 follow, and more details are provided within subsequent pages of this report:

- *Missing Data, Edit, and Imputation*: (1) researched modeling approaches for using administrative records in lieu of Decennial Census field visits in light of forthcoming design decisions; (2) investigated the feasibility of using third party ("big") data from NPD Group to supplement retail sales survey estimates; (3) implemented the model-based procedure Sequential Regression Multivariate Imputation (SRMI) method for missing product data in the Economic Census; and (4) initiated and supported the first phase of the integration of *Tea* to post process American Community Survey data.

- *Record Linkage:* applied and made updates to record linkage software.

- *Small Area Estimation*: (1) developed a small area unit level model which uses a large scale sample survey rather than a population roster for prediction; and (2) created a framework to test the significance of year-to-year changes in small area estimates.

- *Survey Sampling-Estimation and Modeling*: (1) analyzed data from the Commodity Flow Survey to inform decisions by the Bureau of Transportation Statistics on conversion to an all- or primarily-electronic data collection strategy, summarizing response-rates and data-quality by response mode and estimating costs; (2) compared the quality of 2010 Census Nonresponse Follow-up proxy responses and administrative records for the same housing units in the same timeframe using the results of the 2010 Census Coverage Measurement Program; and (3) derived new exact and simple optimal algorithms for allocating the sample considering costs and statement of desired precision.

- *Time Series and Seasonal Adjustment*: (1) completed new methodology of extreme-value adjustment for seasonal adjustment, with application to New Zealand agricultural data; (2) developed an R user-interface for X-13ARIMA-SEATS, allowing for greater usability and communicability of seasonal adjustments; and (3) developed new estimation methods for vector time series models, allowing for parameter constraints.

- *Experimentation and Statistical Modeling:* developed code to carry out exhaustive variable selection for the MAF Error Model; ran selection process on Decennial Statistical Studies Division (DSSD) address canvassing database plus six supplemental data sources; considered two-way interactions for all main effects; and used randomized quantile residuals and other evaluations to study the selected zero-inflated negative binomial model.

- *Simulation and Statistical Modeling:* (1) developed new model-based methodology for analyzing singly imputed synthetic data under exponential and univariate normal models and developed similar methods for multivariate normal data assuming all variables are synthesized, and multiple linear regression models when the response variable is synthesized; (2) developed new methodology for using noise multiplication as an alternative to top coding for statistical disclosure control under a log-normal model where the log-scale mean is modeled as a linear regression; and (3) created an initial artificial population for simulating the Monthly Wholesale Trade Survey over a two-year period and used it to evaluate imputation and estimation methodology.

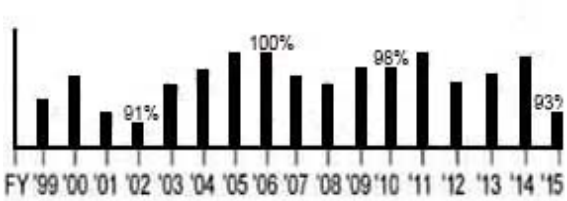
- *SUMMER AT CENSUS:* Sponsored, with divisions around the Census Bureau, scholarly, short-term visits by 30 researchers/leaders who collaborated extensively with us and presented seminars on their research. For a list of the 2015 *SUMMER AT CENSUS* scholars, see *http://www.census.gov/research/summer_at_census/*.

How Did We[1] Do...

For the 17th year, we received feedback from our sponsors. Near the end of fiscal year 2015, our efforts on 28 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 28 questionnaires were obtained with the following results (The graph associated with each measure shows the performance measure over the last 17 fiscal years):

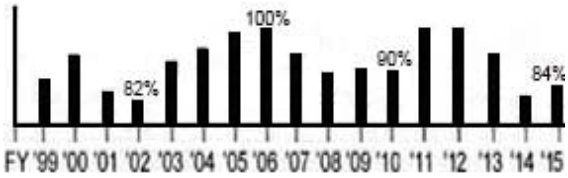
Measure 1. Overall, Work Met Expectations

Percent of FY2015 Program Sponsored Projects/Subprojects where sponsors reported that overall work met their expectations (agree or strongly agree) (26 out of 28) ……..……..... 93%



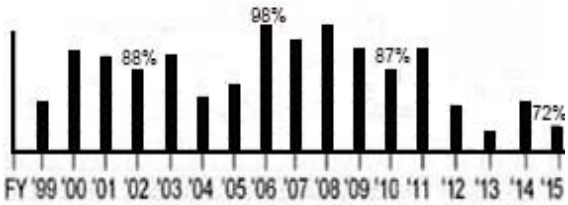
Measure 2. Established Major Deadlines Met

Percent of FY2015 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met (16 out of 19 responses) ……....……... 84%



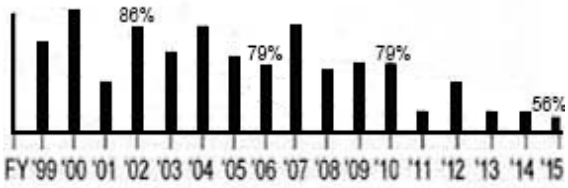
Measure 3a. At Least One Improved Method, Developed Technique , Solution, or New Insight

Percent of FY2015 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight (18 out of 25 responses) …… 72%



Measure 3b. Plans for Implementation

Of these FY2015 Program Sponsored Projects/Subprojects reporting at least one improved method, technique developed, solution, or new insight, the percent with plans for implementation (10 out of 18 responses) …….. 56%



Measure 4. Predict Cost Efficiencies

Number of FY2015 Program Sponsored Projects/Subprojects reporting at least one "predicted cost efficiency" ……..…… 6

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

Measure 5. Journal Articles, Publications

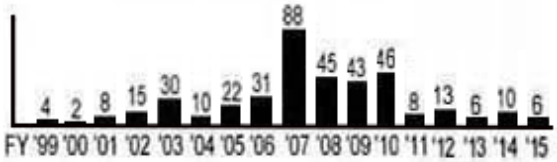
Number of peer reviewed journal publications documenting research that appeared (14) or were accepted (7) in FY2015 …….... 21

Measure 6. Proceedings, Publications

Number of proceedings publications documenting research that appeared in FY2015 ……. 13

Measure 7. Center Research Reports/Studies, Publications

Number of center research reports/studies publications documenting research that appeared in FY2015 …….. 6



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 COVERAGE IMPROVEMENT DESIGN (Decennial Project 6410A00-D03)

1.2 NON RESPONSE FOLLOW-UP DESIGN AND OPERATIONS (Decennial Project 6410A00-D22)

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: Inactive.

Staff: William Winkler (x34729), William Yancey

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 2015, staff started discussions on transitioning from a purely model-based synthetic estimate of net coverage to small area shrinkage (composite) type estimator for the 2020 Census. For areas that are not in the census coverage measurement survey sample, the estimator will remain the same (purely synthetic).

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 combined with the use of administrative records (AR) files in different ways in order to reduce the cost of the operation while maintaining data quality. Regardless of the contact strategy, the question arises as to 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: Staff collaborated with Decennial Statistical Studies Division staff to compare the quality of 2010 Census NRFU enumerations to the ARs available for the same housing units. The study uses Internal Revenue Service and Medicare records from all of 2010 at the addresses in the CCM sample block clusters. The ARs are linked to both the NRFU enumerations (E sample) and the CCM population sample (P sample) interviews conducted independently from the census. When a linking E or P sample record is found for an AR, the residence status assigned by CCM to the record is applied to the AR. This determination of whether the person lived at the CCM sample address on Census Day enables evaluating the quality of the ARs. Staff presented results at the 2016 Joint Statistical Meetings and submitted a paper to its proceedings. Staff distributed a more detailed draft report for internal Census Bureau for review.

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: Inactive.

Staff: William E. Winkler (x34729), William E. Yancey, Tom Mule (DSSD), Lynn Imel (DSSD), Mary Layne (CARRA)

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 2015, staff continued to use data files of 2010 Nonresponse Follow-up (NRFU) IDs for Maricopa County to produce cross tabulations of potential covariates with household size for two main cases: occupied NRFU housing units (HUs) and those NRFU units where the 2009 IRS 1040 return was present and the undeliverable as addressed (UAA) flag was blank. Staff began running models (primarily logistic regression) with 2010 Census unedited file (CUF) household size as the dependent variable, mostly in the above two cases. The only models where the predicted distribution of household size did not seriously distort the actual household size distribution were where a separate logistic regression was fit for each value of IRS 1040 count (large IRS 1040 counts

were combined into a single model). The models did not appear to perform much better than simply using the IRS 1040 count. Staff sent a draft overview of the Maricopa results to a subgroup of the Administrative Records Modeling Team.

Staff also used national data files of 2010 NRFU IDs to produce crosstabulations and correlations of potential covariates with CUF household size. The crosstabulations and correlations focused on potential covariates that were not included in earlier data files and on the subset of NRFU IDs where the 2009 IRS 1040 return was present and the UAA flag was blank. Staff began running national models (primarily logistic regression) with 2010 CUF household size as the dependent variable. The models focused on two cases: NRFU IDs where the UAA flag was blank and the 2009 IRS 1040 was present, and NRFU IDs where the UAA flag was blank and the 2009 IRS 1040 was absent but the 2008 IRS 1040 was present. The only models where the predicted distribution of household size did not seriously distort the actual household size distribution were models where a separate logistic regression was fit for each value of IRS 1040 count (large IRS 1040 counts were combined into a single model). The models did not appear to perform much better than simply using the IRS 1040 count. The models for the case where the 2009 IRS 1040 was absent but the 2008 IRS 1040 was present had less predictive power than the models for the case where the 2009 IRS 1040 was present. Staff sent selected output from the national models along with explanatory text to a subgroup of the Administrative Records Modeling Team.

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 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 the first half of FY 2015, staff researched to determine the strategy for using administrative records in the 2015 Census test. As part of the administrative records modeling team, staff compared the predictive power of a variety of models and researched how to combine the models into one determination for removing housing units from the NRFU workload. Focusing on the analysis of 2010 Census data from Maricopa County, AZ (the location of the 2015 Census test), staff presented these ideas and preliminary findings at the "2020 Research Workshop: Administrative Records" organized by CSRM and Decennial Statistical Studies Division (DSSD) chiefs.

Staff investigated linear programming techniques with the objective of maximizing the use of administrative records subject to varying measures of quality of the administrative records. This approach was implemented in the 2015 Census test. Staff also observed Census test NRFU enumerators in Mesa, Arizona and completed a field observation report documenting the experience, including issues encountered in the field and recommendations for improving procedures. During the second half of FY 2015, staff presented national-level results of the linear programming methodology in a 2015 Joint Statistical Meetings (JSM) invited session on "Utilizing Administrative Records and Adaptive Design in the 2020 Census." Staff submitted a *JSM Proceedings* paper in conjunction with this talk. Throughout FY 2015, staff continued research on the comparison of classification methods for a person-place model for administrative records usage and submitted a paper on the topic to a peer-reviewed journal.

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: 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 using 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), and the other using data from the Recall Bias Study, which was part of the 2010 Census Evaluation and Experiments Program. Draft papers containing the methodology and results regarding recall error underwent review. Staff continues

to improve the manuscripts through addressing the comments received from reviewers. In addition, an invited paper regarding lessons learned about evaluating survey data with administrative records files was presented at the 2015 Total Survey Error Conference. The paper will appear in a monograph of the conference invited papers.

Staff: Mary Mulry (x31759)

H. Census Cost-Quality Tradeoff Assessment
Description: This project assesses the tradeoff between cost and quality for the key design decisions for the 2020 Decennial Census. The Cost-Quality team coordinates this project.

Highlights: During FY 2015, staff reviewed assorted memoranda related to the cost and quality effects of using administrative records in Nonresponse Follow-up (NRFU). Staff used these memoranda to produce spreadsheets of quality indicators for alternative scenarios for using administrative records in NRFU in the 2010 Decennial Census and the 2014 Maricopa Census Test. The spreadsheets and explanatory text were sent to the chair of the team as information for a planned memorandum from the team outlining cost and quality scenarios for key elements of the 2020 Decennial Census.

Staff: Michael Ikeda (x31756)

1.3 ADDRESS LIST QUALITY MEASUREMENT (Decennial Project 6310A00-C02)

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 2015, staff refined the zero-inflated count models developed to characterize and predict MAF errors for census tabulation blocks. Staff carried out stepwise variable selection on the main 2009 address canvassing database prepared by the Decennial Statistical Studies Division (DSSD) along with six supplemental data sources to select predictors and two-way interactions and to determine which are most beneficial to the model. Staff found that 2009 add counts for some blocks were not explained by available covariates. This occurred in a very small proportion of blocks in the U.S., but many of these blocks had moderate to large add counts. A report based on this work is being completed. Staff worked with an external collaborator to revise a manuscript on his earlier MAF modeling work, which has been accepted for publication. Staff supplied DSSD personnel with models for the data analysis in their report on the 2015 Address Validation Test. Staff began work on block-level count models with spatial random effects, e.g. using a conditional autoregressive structure in a Bayesian framework, to capture some of the variability in observed counts which has not been explained by earlier models.

Staff: Andrew Raim (x37894), Marissa Gargano, Kimberly Sellers, Deborah Fenstermaker (DSSD), Laura Ferreira (DSSD), Krista Heim (DSSD), Derek Young (University of Kentucky)

B. Census Enterprise Microsimulator
Description: The Census Enterprise Model (CEM) is a microsimulation tool developed by The MITRE Corporation jointly with the Census Bureau. Given a list of households, a decision tree describing possible census outcomes, and a set of probabilities corresponding to those outcomes, the CEM provides simulated realizations of a census. These realizations can be used to compare cost and quality among various competing strategies being considered for the 2020 Decennial Census.

Highlights: Staff began formulating a Bayesian approach to decennial census microsimulation. Under this approach, the unobserved binary outcomes of whether or not households respond to the 2020 Census are simulated from a posterior predictive distribution. Staff also prepared an internal document clarifying the validity of the normal approximation (based on the central limit theorem) in a specific non-standard situation.

Staff: Martin Klein (x37856), Andrew Raim, Deborah Fenstermaker (DSSD), David Brown (CES)

C. 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 2015, the TRMAC team worked to identify areas in which the MAF was not complete or accurate. The product of this activity was a Block Tracking Database (BTD) for tracking the status of the 11.1 million tabulation blocks in preparation for in-office canvassing and in-field canvassing. During the first and second quarters, staff developed and maintained the BTD.

Geography Division (GEO) and Center for Statistical and Methodology (CSRM) developers and contractors began to develop the BARCA using Java code, SQL script, and Agile development methods. The BARCA incorporates the Census Bureau's TIGERWeb application, as a web-mapping service (WMS), to display TIGER features and reference layers. The BARCA in-office canvassing has started in production and the Block Tracking Database (BTD) is tracking the status of the 11.1 million tabulation blocks in preparation for in-field canvassing. The initial in-office canvassing is estimated to take two years to complete.

Staff: Tom Petkunas (x33216)

1.4 CODING, EDITING, AND IMPUTATION STUDY (Decennial Project 6510A00-E01)

A. Software Development (*Tea*)

Description: Here staff report applications of *Tea* software to the coming 2020 Census. *Tea* is software for the editing, imputation, and disclosure avoidance of surveys. It is intended to be easily reconfigured for new surveys. By putting all of these processes in one package, staff can guarantee that all imputations can pass edit requirements, and staff can use advanced imputation techniques to synthesize data that would otherwise fail disclosure avoidance requirements. *Tea* is based on R and several packages for data processing, and it is documented according to professional standards.

Highlights: Staff continued testing of *Tea*'s edit system in the context of the American Community Survey, rewriting key sections of the existing system using *Tea*. Staff also continued testing of *Tea*'s imputation system in the context of the Decennial Census, executing a full count and characteristic imputation of the 2014 test census using *Tea*. Staff rewrote about a third of the code base to better accommodate deterministic edits and edits with no subsequent imputation. Staff presented *Tea* to interested parties in the Decennial Statistical Studies Division (DSSD), the Decennial Management Division (DMD), the Population Division (POP), and the designers of the upcoming CARDS system. Staff developed a public tutorial to help users outside the Census Bureau learn to use *Tea*.

Staff improved the handling of deterministic edits of imputed data, the handling of edits with no imputation, and the expectation-maximization algorithm intended for use in imputation using administrative records. Staff also implemented a system to auto-calculate the structure of dependencies across data tables. Staff extended public documentation to about 25pp, which is available as http://b-k.github.io/tea-tutorial/ or via http://dx.doi.org/10.5281/zenodo.19931.

Staff: Rolando Rodriguez (x31816), Ben Klemens

B. Software Analysis and Evaluation

Description: This project will compare competing imputation methods for the 2020 Decennial Census. Staff will establish testing procedures for the comparison and will produce statistical and graphical output to inform any production-level decisions. The current donor-based imputation method will be tested along with numerous other methods, both from in-house software and from external sources (where feasible). Coordination with production divisions will help ensure that the procedures meet all the necessary production criteria.

Highlights: Staff has made numerous improvements to the codebase of *Tea*, particularly in regards to editing. Research continues into imputation models for use in a collection operation involving reduced non-response follow-up. This includes the use of administrative records for substitution and as model covariates.

Staff: Rolando Rodriguez (x31816), Ben Klemens, Yves Thibaudeau

1.5 DEMOGRAPHIC ANALYSIS (Decennial Project 6710A00-G02)

A. Demographic Analysis

Description: As part of the planning process for the 2020 Census, this project reevaluates the methods used for intercensal population estimates and their variance.

Highlights: Inactive.

Staff: Ben Klemens

1.6 PRIVACY AND CONFIDENTIALITY STUDY (Decennial Project 6810A00-H11)

A. Privacy and Confidentiality for the 2020 Census
Description: This project undertakes research to understand privacy and confidentiality concerns related to methods of contact, response, and administrative records use which are under consideration for the 2020 Census. Methods of contact and response under consideration include internet alternatives such as social networking, email, and text messages. The project objectives are to determine privacy and confidentiality concerns related to these methods, and to identify a strategy to address the concerns.

Highlights: Inactive.

Staff: Martin Klein (x37856)

1.7 MATCHING PROCESS PROGRAM DESIGN AND EXECUTION (Decennial Project 6810A00-H09)

A. 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 2015, staff prepared a draft executive summary of initial results from a matching of the 2010 Census unedited file (CUF) against itself. Staff sent the draft executive summary and a more detailed draft overview to the Division Chief of Decennial Statistical Studies Division (DSSD). Staff reviewed and provided comments on the draft memoranda "Person Computer Matching for Standardization of the String Comparator for Person Matching" and "Decennial Person Computer Matching Research for Modifications to Invalid Name Lists and Invalid Name Cleanup Software Functions." The memoranda are being released as part of the *DSSD 2020 Decennial R&T Memorandum Series.*

Staff is also working on revising the parameter files and related sections of the CSRM Matcher. The CSRM Matcher performs one-to-one matching for relatively small data sets. Finally, staff joined the new Automated Matching Team (the replacement for the old Matching Process Improvement Team) and participated in the planning and preparation process for the team. As part of this process, staff sent the draft executive summary and draft overview mentioned above to the chair of the Automated Matching Team. The Automated Matching Team has currently suspended activities while the Decennial area considers priorities for the team.

Staff: Michael Ikeda (x31756), Ned Porter, Bill Winkler, Bill Yancey

1.8 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 6385A70)

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: Staff revised algorithms for generating custom multi-year estimates in order to separate the effects of sampling and non-sampling uncertainty and to obtain formulas for parameter estimates, thereby obviating the need to do optimization. Staff met with clients from Veterans Administration to discuss preparing the method for production on 30,000 time series.

Staff: Tucker McElroy (x33227)

B. ACS Imputation Research and Development
Description: The American Community Survey (ACS) process of editing and post-edit data-review is currently time and labor intensive. It involves repeatedly submitting an entire collection year of micro-data to an edit- enforcement program (SAS software). After each pass through the edit-enforcement program, a labor-intensive review process is conducted by a staff of analysts to identify inconsistencies and quality problems remaining in the micro-data. Before the data are ready for public release, they have at least three passes through the edit-enforcement program and three review processes by the analysts, taking upward of three months. The objective of this project is to experiment with a different strategy for editing—while keeping the same edit rules—and to assess if the new strategy can reduce the number of passes through the edit process and the duration of the review process.

Highlights: Staff further integrated *Tea* in the ACS survey post-processing environment. Staff met repeatedly with staff from the Population Division (POP) and the Social, Economic, and Housing Statistics Division (SHESD) to identify a set of edit rules to establish a "proof of concept" to illustrate the basic structure and functionality of *Tea*. Staff translated the ACS edits into "checks" in the syntax of *Tea. Tea* follows the logic of the checks and enforces all the logic

rules associated with the checks. Staff provided support by editing the "REAP" so that *Tea* remains a project in good standing the decennial area.

Staff: Yves Thibaudeau (x31706), Ben Klemens, Rolando Rodriguez

C. 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 2015, staff completed the statistical design for an August 2015 Pilot Experiment on intervening to remove from Field Representative (FR) follow-up all ACS sampled units for which a cumulative respondent-burden score exceeded a specific threshold. Staff also participated in discussions concerning the implementation of the pilot study and debriefing of FR and Field Supervisory participants in the pilot; analyzed the statistical results from the Pilot data; and began drafting the final research report summarizing the findings and conclusions.

Early in FY 2015, staff efforts in cooperation with the CAPI Max Burden Team included comparison among many competing proposed policies for removing cases from FR workloads based on 2012 ACS data, and on choosing a specific case-removal policy based on a cumulative Burden Score designed by the Team. This policy of curtailed FR case-follow-up is due to implemented nationally in the spring of 2016. The statistical design for the pilot consisted of 12 selected Survey Supervisor Field Areas (SSFAs) (from matched pairs within Regional Offices, out of 48 national SSFAs) and block-randomized Field Supervisory Areas within selected SSFAs assigned to one of three treatment groups differing by whether cases would be removed from FRs or by whether individual FRs would be aware of the Burden Scores on which case removals were based. The pilot was conducted throughout August 2015, after which staff linked the available data sources—Unified Tracking System data on FR transmissions and on FR reporting in the Contact History Instrument (CHI), and Cost and Response Management Networks (CARMN) data on FR payroll submissions—and analyzed the data to compare metrics of ACS response rate, cost, and respondent burden across treatment groups by contrast with ACS SSFAs not selected to participate in the pilot. Findings and statistical methodology are being drafted into a final report to be reviewed and included in the ACS research report series in late calendar 2015.

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

D. 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 1871000).

Highlights:

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

Staff: Tommy Wright (x31702), Martin Klein, Brett Moran, Derrick Simmons, Nathan Yau

E. Confidence Intervals for Proportions in ACS Data

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

F. Mode-Based Imputation in ACS

Description: It is well known that item nonresponse differs markedly between the different modes of data collection within the American Community Survey (ACS), yet current ACS practice is to perform item imputations via hot-deck methods pooling all ACS modes together. This project investigates the impact of doing ACS item-imputation separately within modes (Mail, CATI, CAPI) using the 2012 year of ACS data. It does this by developing a model-based mass-imputation approach to imputation using categorical control variables as similar as possible to those used in current hot-deck imputation and by comparing the impact on ACS estimates for selected survey variables if the mass imputation is done ignoring mode vs. being done within cells cross-classified by mode.

Highlights: During FY 2015, staff completed data analysis and the writing and revision of an ACS research report on the possible impact of mode-based imputation in ACS as viewed through comparisons between model-based imputations that did take response mode explicitly into account versus those that did not. The report, which appears in the *ACS Research Memorandum* series (ACS15-RER-07), found little impact of mode in addition to other natural control variables in single-item model-based imputation, but found that the relationships between observed and imputed items within single case records are often dramatically different from those of jointly observed items. Additional outcomes were proposed to further research on (1) characterizing differences between the hot-deck-donor population and

the general population for ACS survey variables not only for the missing-data units but for all ACS-sampled units and (2) creating new variables for ACS model-based item imputation based on the ACS hot-deck item allocations.

Staff: Eric Slud (x34991), Ryan Janicki

G. Voting Rights Section 203 Model Evaluation and Enhancements Towards Mid-Decadal Determinations

Description: Section 203 of the *Voting Rights Act* asks for determinations relating to limited English proficiency and limited education of specified small domains (race and ethnicity groups) for small areas such as counties or minor civil divisions (MCDs). Section 203 seeks to determine whether or not small areas must provide voting materials in languages other than English. Previous research undertaken provided a small area model-based estimate derived from American Community Survey (ACS) 5-year data and 2010 Census data, which provides smaller estimated variances than ACS design-based estimates in many cases. Research and groundwork into the production mid-decade determination is ongoing.

Highlights: During FY 2015, staff evaluated aspects of the model applied during FY 2011 for producing determinations under Section 203. The focus was to start with creating simulations investigating baseline fundamentals relation to small area estimation and hierarchical models to build to the point of incorporating aspects of survey design behavior. Staff identified areas in which estimation methodology can be improved. This is characterized by two different approaches: 1) by eliminating all jurisdictions by language minority group combinations whose design estimates do not come close to the thresholds of Section 203 inclusion/exclusion (measured by the number of standard errors from the threshold) and then formulating a model, and 2) by continuing the FY 2011 model with the notion that improvements are made to the model as well as estimation method adjustments made to 2010 Census as some feel it should not be relied on for detailed race and ethnicity information to make applications using 2014 American Community Survey Data. Staff has made progression towards these goals by using 2012 American Community Survey data in order to refine methods and computational programs.

During this time, CSRM staff held consultations with various Census Bureau staff to understand what the *Voting Rights Act* requires of the Census Bureau. These consultations also summarized previously used methods for making determinations under the act and current progression of research, study, and evaluation. Documentation of progression and research were often included in these consultations.

All work is ongoing with the intention to end research and work towards production in FY 2016 with a data release in FY 2017.

Staff: Patrick Joyce (x36793), Eric Slud, Tommy Wright, Tom Louis (ADRM)

1.9 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 progress. This project will continue with Demographic Statistical Methods Division (DSMD) and CSRM staff 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: Regression survey estimators based on nonresponse adjustments are generally expressed through estimating equations in which survey weights are adjusted for nonresponse through estimated propensities. The propensities are generally functions of survey variables observed for all sampled individuals, yet the validity of the estimating equation estimators depends on an assumption that the outcomes and response indicators are conditionally independent, something that would be much more plausible if the propensities were allowed to depend also on variables involving demographics observed at the individual level only for survey-responders. This project is devoted to theoretical research on methods of weighted estimating equations applicable to surveys in which propensities may depend on covariates observed only for responders, which are illustrated on American Community Survey data.

Highlights: During FY 2015**,** staff developed propensity models and estimating equations applicable when the population joint distributions are either known exactly or well-described by a joint density modeled from external data such as national Population Estimates. A

2015 Joint Statistical Meetings paper and talk and a NISS Workshop (November 2015) talk were prepared from this material. The new results concern the form of optimal weight estimators derived from propensity models and the form of model-assisted (double-robust) regression which together minimize the variance of the estimate of a population total or mean. Research is continuing into alternative forms of estimating equations to be applied to ACS data for estimating population totals.

Staff: Eric Slud (x34991)

1.10 NATIONAL SURVEY OF COLLEGE GRADUATES (Demographic Project 7491013)

Description: Staff will develop models that predict response mode prior to and throughout the NSCG data-collection period. The primary objective is to determine whether we can assign sample members to their preferred mode as quickly as possible to shorten the data-collection period while maintaining response rates and representativeness.

Highlights: After examining a variety of models to predict response mode and time to response by mode, staff determined a set of potential experiments for the next round of the National Survey of College Graduates. These experiments include starting some sample members in mail or Computer-Assisted Telephone Interviewing (CATI) at the beginning of data collection (rather than by the default Internet mode), offering some sample members more than one initial mode response option, and offering some sample members mail or CATI early.

Staff also designed 11 interventions to take place every two weeks during the 2015 National Survey of College Graduates data-collection period, beginning in week four. These interventions use the models developed earlier in the year to generate real-time response propensity estimates to determine when to offer response mode options to sample members, or to suspend follow-up in a particular mode.

Staff: Chandra Erdman, Stephanie Coffey (DSMD)

1.11 DEMOGRAPHIC SURVEYS DIVISION (DSD) SPECIAL PROJECTS (Demographic Project 0906/1442)

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 2015, staff proposed to perform simulated attacks on the Microdata Analysis System (MAS). By making numerous queries to the MAS and mining the results, staff attempt to identify individuals that have participated in the American Community Survey, and possibly follow-up with similar attempts with other demographic sample surveys. Necessary data access must be granted in order to perform these simulations. Staff continued to attack the Public Use File of the American Community Survey.

Also during FY 2015, staff attacked the Public Use File of the American Housing Survey from years 2009-2013 to see if it can identify any participant or house that took part in the survey. Staff concentrated attacks on the 2009 files on houses whose last purchased price was not divisible by 10 and are more vulnerable to reidentification. Staff has written software for integrating the data into attackable files that performs extractions of necessary data and scrapes popular real estate websites to see if one can locate a house. The American Housing Survey asks questions about one's relationship with his/her neighbors, so any identification would be considered a disclosure. Staff introduced outside commercial data for attacking public use files of 2009-2013 American Housing Survey. Phase 1 of the study includes the merging the 2014 CoreLogic commercial data. From this phase, staff can identify and rank variables that are vulnerable for reidentification.

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

1.12 NATIONAL CRIME VICTIMIZATION SURVEY (Demographic Project 7523013)

A. National Crime Victimization Survey (NCVS): Contact Response Propensities and Adaptive Design Research

Description: A cross-directorate team has been established to develop and evaluate models that predict near-real-time response propensities in the National Crime and Victimization Survey. The team will monitor daily propensities along with actual survey response indicators and survey outcomes to (1) evaluate model accuracy and determine whether the models need refinement; (2) investigate relationships between response propensity and key survey variables; and (3) determine how the propensities may be used to manage fieldwork.

Highlights: During the first half of FY 2015, staff modeled attempt-level response propensities with block group-level characteristics, sample frame data, and paradata from prior contact attempts. Staff in the Center for Survey Measurement drafted a paper describing the steps to identify the best predictors of response rate, model design choices that affect the ability to predict response propensities, how model and variable choices affect the stability of predicted propensity distributions from month to month, and the predictive power of the best model. The results were presented at an internal seminar.

During the second half of FY 2015, staff completed an internal comprehensive draft report documenting work on NCVS response propensity modeling and methodology to provide additional information to serve as a guideline for Field Representatives (FRs) to create intervention for NCVS fieldwork for cases with low propensities. Staff is currently developing this draft report. The final findings and the report series will be presented to the sponsor.

Staff would like to show how this response propensity methodology can provide potential intervention strategy for survey efforts to increase the NCVS response rates, and how more interview cases can results in completions. The response propensity modeling is currently being developed to help reverse the NCVS declining response rates, to control survey errors and cost. The revised estimated propensities from propensity models can be used to strategically direct FRs efforts to improve NCVS survey response rates, to improve data quality, and to reduce costs.

Staff: Isaac Dompreh (x36801), Joe Schafer, James B. Lawrence (CSM), Peter Miller (ADRM), Meagan M. Meuchel (ADDP)

1.13 POPULATION DIVISION PROJECTS (Demographic Project TBA)

A. Population Projections

Description: This project provides methodology and software to generate long-term forecasts for fertility, mortality, and migration data using vector time series techniques.

Highlights: Inactive.

Staff: Tucker McElroy (x33227), Osbert Pang, William Bell (ADRM)

B. 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: During FY 2015, staff traveled to Dhaka, Bangladesh and conducted a week-long (May 17-21, 2015) Workshop on Sampling at the well-known public health research organization International Center for Diarrheal Disease Research, Bangladesh (ICDDR-B). 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 estimation of population parameters from sample surveys as a function of sample design, weighting procedures, the computation of sampling errors of sample estimates, and the making of inferences from the sample to the population. The twelve workshop participants were mostly senior researchers with graduate backgrounds in public health and statistics whose work involved many health-related sample surveys of the Bangladeshi people. Three participants were associated with the University of Dhaka (one faculty member and two graduate students in Statistics). The workshop was sponsored by ICDDR-B, USAID-Bangladesh, and the U.S. Bureau of the Census.

Staff: Tommy Wright (x31702), Michael Leibert

1.14 SOCIAL, ECONOMIC, AND HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165015)

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 2015, staff continued to explore ways of borrowing information over time for inference about county-level estimates of children in poverty and expanded the results to include three different years of estimation (2010, 2011, 2012) to ensure patterns are consistent among years. The models studied included time series and bivariate extensions to the Binomial-Logit Normal (BLN) Model on ACS direct estimates of

counts and to the linear Fay-Herriot model applied to ACS estimates of rates. These were implemented through a hierarchical Bayes approach. A generalized variance function (GVF) was used to stabilize sampling variance estimates.

For both the BLN and Fay-Herriot models, staff contrasted predictions and prediction standard errors from the bivariate and time series extensions to those of the univariate version. The bivariate model appeared to perform best based on these criteria. Staff compared these results to theoretical results regarding the behavior expected from using these alternative approaches to borrowing information over time, under simplifying assumptions. Staff also contrasted variations of the models, including a heteroscedastic version, with different model variances among dependent variables as well as a homoscedastic version, where these variances are assumed equal. Staff examined the extent to which the posterior predictive distributions of the BLN model showed evidence of skewness.

The bivariate versions of both the BLN and Fay-Herriot models were contrasted to the SAIPE production model. In particular, staff compared the predictions, prediction standard errors, and the widths of confidence intervals to those of the production model. Staff began exploring whether the choice of link function, and its implications on the relationship of model variances and means, has an impact on prediction standard errors. Staff also studied properties of a hierarchical Bayes multivariate Fay-Herriot model, including proving the propriety of the posterior distribution under a class of priors. A bivariate multivariate measurement error model was applied to SAIPE data for inference on school-aged children in poverty for counties.

Staff researched diagnostics for small area estimation models, and applied them to the BLN model. The diagnostics applied involved the evaluation of performance of synthetic estimators when sample sizes are large, as well as comparisons of raking factors. Staff also generalized the BLN and Fay-Herriot Models to have model errors that follow a heteroskedastic AR(5) structure, and applied the models to SAIPE data. Staff contrasted posterior predictions and standard errors of the AR(5) extensions to those of analogous AR(1) and bivariate SAE models. Staff presented this work in an invited session on poverty mapping on the Fourth Italian Conference on Survey Methodology in Rome, Italy. Staff wrote a paper that was accepted for publication.

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.

Bayesian Benchmarking of Estimates from Distinct Geographic Models
Highlights: A paper documenting work on this project was written and submitted for publication during FY 2014. Work on this project is now complete.

Comparing Small Area Estimates Over Time
Highlights: During FY 2015, staff worked on the problem of estimating year-to-year change of the proportion with health insurance in different small areas and quantifying how different the change is from zero. For the linear model, a comparison of predictions using a bivariate normal distribution on model errors was made with predictions from a model using an AR1 structure. Credible intervals were constructed using each method. A paper documenting this work was completed. Preliminary work was done on extensions to multivariate, nonlinear models, with special focus on multivariate Beta regression models. Different transformations of predictions to achieve approximately normally distributed residuals were constructed, and estimators of the model covariance based on these transformed residuals were derived.

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 2015, staff finalized the simulated datasets and fitted the generalized variance functions (GVF) to supply a model-based sampling variance estimates for the area level aggregates. Staff also developed and implemented the evaluation framework. As a test, the direct estimates with three different versions of sampling error variance estimates (direct, GVF and the true sampling variance). The GVF showed slightly more bias than the direct estimates of the sampling variance, but the confidence intervals based on the GVF estimate of the sampling variance performed better than the direct design-based method (ACS replicate weights) 86.8% to 80.0% (for 90%

confidence intervals). The intervals based on the true sampling variance were conservative, 91.4% coverage, most likely due to the poor normal approximation at these smaller sample sizes.

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

1.15 ECONOMIC STATISTICAL COLLECTION (Economic Project 1183001)

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: In the previous phase of this project, staff found that the current survey estimators (Horvitz-Thompson (HT) and Random Groups) do not perform as large sample theory suggests they should. Based on these findings, staff outlined an alternative to the HT based procedures for the MWTS. The alternative procedure is model-based and uses Bayesian methodology in a finite population setting. The current phase of research is mid-stream. Staff have proposed deriving estimates using (1) a parametric model (multivariate normal) to preserve key intervariable relationships, (2) nonparametric univariate density estimates to preserve marginal distributions for each variable, and (3) a copula model to link (1) and (2). Staff presented these ideas at the 2015 Joint Statistical Meetings and prepared a paper based on this work for the conference proceedings.

Staff are also working to improve upon the simulated wholesale trade population. The improvements in the artificial population focus on relaxing certain simplifying assumptions; in particular, the assumption that all sampled businesses operate in only one industry class.

Staff: Martin Klein (x37856), Joe Schafer, Joanna Fane Lineback (ADEP), 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 2015, staff members joined Economic Directorate staff on a new "Retail Big Data Team." The long-term purpose of this team is to develop a team of experts with knowledge of big data tools and resources. During the first half of FY 2015, the team evaluated data from a market research company named NPD Group. In response to a request for proposal (RFP), NPD provided aggregated scanner data for two types of products: auto and jewelry/watches. CSRM staff assessed the quality of the NPD data (e.g. the completeness of the NPD data series for specific levels of geography) and developed strategies for comparison to retail survey data and Economic Census data, achieving an end goal of specifying a set of recommendations about using such data to supplement current Census retail data. CSRM staff co-authored the final report entitled "The Feasibility of Using NPD Data to Supplement Retail Trade Data."

At the end of FY 2015, CSRM staff initiated involvement with a sub-team of the "Retail Big Data Team" (in collaboration with researchers at the Bureau of Economic Analysis (BEA)) tasked with evaluating the quality of aggregated MasterCard data. Staff completed preliminary analysis of completeness of the data series and comparisons to Annual Retail Trade Sales estimates. Additionally, as part of the retail big data team, staff also participated in the big data working group – a group formed to share progress from big data projects underway in the Economic Directorate.

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

1.16 ECONOMIC CENSUS/SURVEY ENGINEERING: TIME SERIES RESEARCH (Economic Project 2220A10)

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: Staff provided support for seasonal adjustment and software support for users within and outside the Census Bureau, including IMC Asset Management, International Paper, M&T Bank, Macrobond, Newsignature, OPUS, PayNet Inc., PennyMac, Putnam Investments, RS Metrics, SAS, Vancouver International Airport, Virginia Bioinformatics Institute, White Pine Investment Management, Administrative Office of the U.S. Courts, Bureau of Labor Statistics, Colorado Department of Labor and Employment, Cuyahoga County Government, Federal Reserve Bank of Minneapolis, Michigan Department of Licensing and Regulatory Affairs, National Center for Health Statistics, New Jersey Department of the Treasury, U. S. Department of Transportation, Australian Bureau of Statistics, Austrian Bureau of Statistics, Central Bureau of Statistics (Israel), Federal Statistical Office (Switzerland), INSEE (France), Israel Tax Authority, UK Office for National Statistics, KOSTAT (South Korea), Statistics Canada, Statistics New Zealand, Statistics Norway, Statistics South Africa, Turkish Statistical Institute, Bank of Korea, Bundesbank, Central Bank of Venezuela, International Monetary Fund, SCHB (Mexico), Swiss National Bank, Chiba University (Japan), Columbia University, Katholische Universität Eichstätt-Ingolstadt, Nankai University, Northwestern University, Paris University, Ramon y Cajal University Hospital, Stony Brook University, University at Albany, University of Basel, University of California at Berkeley, and University of Southampton.

Staff discussed seasonal adjustment issues with analysts from the Korean Institute of Labor, along with the Time Series Methods Staff from the Economic Directorate. Staff discussed specific modeling and adjustment issues for transportation data with staff from the U.S. Department of Transportation.

Staff organized a visit by Lujuan Chen of the Australian Bureau of Statistics to discuss issues in seasonal adjustment and software. Staff also organized a visit by Christophe Sax, a consultant who has developed the seasonal R package that interfaces with X-13ARIMA-SEATS, to give a talk and have discussions with Census Bureau personnel on future work related to the seasonal package.

Staff discussed issues in seasonal adjustment and moving holiday adjustment with analysts from the Short-Term Service Statistics Division of KOSTAT (South Korea), along with the Time Series Methods Staff from the Economic Directorate. Staff organized a visit by statisticians from TurkStat, the Turkish Statistical Office, to discuss issues in seasonal adjustment, along with staff from the Time Series Methods Staff from the Economic Directorate.

Staff gave a talk for a group of seasonal adjustment experts at the Bureau of Labor Statistics to describe the state of the X-13ARIMA-SEATS project and an R package that interfaces with X-13ARIMA-SEATS.

Staff: Brian Monsell (x31721), James Livsey, Tucker McElroy, Osbert Pang, Anindya Roy, David Findley (Consultant), 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 released an updated version of X-13ARIMA-SEATS (Version 1.1, Build 19), to the Economic Directorate for testing. The Economic Directorate compared adjustments from this version of the software to the last released version of X-13ARIMA-SEATS (Version 1.1, Build 10) and found, in almost all cases, no differences in the adjustments. This version was released to the public.

Staff improved error handling in the AIC automatic testing routines, added an option to compute the QS diagnostic on the logs of the series, and fixed a defect in runs testing. Staff made minor changes to the automatic model identification procedure to ensure a seasonal component is not identified when options are selected that preclude that choice. Staff also made minor changes

in the HTML output of the program to add spacing, and added a check to input routines to test if names input by the user exceed the maximum length defined in the source code.
Staff added two new diagnostics for the adequacy of residuals (Friedman's, Durban Watson) to the software, and added some output to the unified diagnostics file. Tables for outlier adjusted SEATS seasonal adjustment and outlier adjusted SEATS irregular have been added to the program for printing and saving. Staff also implemented the correct stock length of month regressor in the regARIMA modeling module, ensuring the correct forecast parameters are used to produce correct component values and forecasts for SEATS.

Staff updated SEATS seasonal adjustment routines within X-13ARIMA-SEATS and have found very little difference between the SEATS adjustments generated from this version to previous versions of X-13ARIMA-SEATS.

Staff: Brian Monsell (x31721), Osbert Pang, David Findley (Consultant)

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: Staff (a) refined work on preliminary estimation of VARMA models via examination of spectral shrinkage (based off the Hannan-Rissanen method); (b) extended results on using the Frobenius norm to test goodness-of-fit; (c) completed simulation studies of size and power for trading day statistics and the impact of sliding data spans upon the coefficients; and (d) began simulation studies to examine the impact of weather upon seasonal adjustment.

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

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 Seasonal Adjusted Papers site with new and updated research reports and papers.

Staff released updated versions of the Win X-13, X-13-Data, and X-13-Graph programs and released a new interface to the Genhol program, Win Genhol, developed by Nicole Czaplicki of the Time Series and Related Methods Staff of the Economic Directorate. Staff updated and revised documentation files for each of these programs. Later in the year, staff revised the Genhol and Win Genhol programs to ensure they interface more efficiently with X-13ARIMA-SEATS and released these new versions of the software.

Staff developed macro files to generate spec file in UltraEdit from user input. Staff is evaluating an R interface with X-13ARIMA-SEATS and developed several functions to save X-13 files and extract X-13 diagnostics and perform simulation studies using R's facility for parallel computing. Staff compiled a Linux version of the X-13-Graph-Java software, developed by Demetra Lytras of the Time Series and Related Methods Staff of the Economic Directorate.

Staff: Brian Monsell (x31721), Tucker McElroy, James Livsey, Osbert Pang, William R. Bell (ADRM), David Findley (Consultant)

1.17 ECONOMIC CENSUS/SURVEY ENGINEERING: ECONOMIC MISSING DATA/PRODUCT LINE DATA (Economic Project 2220A10)

1.18 ECONOMIC CENSUS/SURVEY ENGINEERING: DEVELOPMENT/SAS (Economic Project 2220A10)

A. 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: Staff contributed to the final report completed in Q1 of FY 2015. As part of a topic-contributed session at the 2015 Joint Statistical Meetings (JSM) in Seattle, staff presented on the two regression-type imputation procedures assessed in the product line imptation project in the talk, "Implementation of Ratio Imputation and Sequential Regression Multivariate Imputation on Economic Census Products" (abstract located at https://www.amstat.org/meetings/jsm/2015/onlineprogram/AbstractDetails.cfm?abstractid=315436). A *JSM Proceedings* paper was submitted on the topic.

Staff was 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 entitled "Using Classification Trees to Recommend Hot Deck Imputation Methods: A Case Study" for the 2015 Federal Committee on Statistical Methodology (FCSM) conference.

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

1.19 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: Based on research questions asked by sponsors, staff developed and carried out a research plan examining the effect of response mode (paper vs. electronic) on the probability of response to the survey, data quality, and cost in the CFS. Staff used a propensity score weighting approach to estimate the response mode effects by identifying the characteristics of establishments associated with response mode. Comparisons of estimated cost were made under different data collection strategies and scenarios. Throughout the year, the intermediate and final results were presented to the survey sponsors, and the results discussed. Finally, staff developed a first draft of a research report summarizing the results and recommendations of the study.

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

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

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

Description: Staff members were asked to develop and carry out statistical methodology to validate the Finance Division's current methodology for estimating Census Bureau's total FY 2014 accruals as of September 30, 2014. The total FY 2014 accruals is the total expenses for work done on contracts or purchase orders that have occurred, but have not been paid. Without contacting every contractor, there is no way to know this total with certainty. So to estimate this total, Finance Division has multiplied the total value of all contracts and purchase orders that have not been paid as of August 31, 2014 by

an average estimate of the ratio of amount paid on contracts/purchase orders to total values of contracts/purchases.

Highlights: During FY 2015 and using data provided by the Finance Division, staff members determined a stratified random sampling plan (with desired sample size) to estimate the total accruals using a 99% confidence interval. The interval from the stratified random sample contained the estimate that the Finance Division had obtained. The complete details are contained in the final September 11, 2015 memo to the Finance Division, "A Sampling Plan to Assess the Census Bureau's Method for Estimating FY 2014 Accruals."

Staff: Tommy Wright (x31702), Carolina Franco, Michael Leibert, Andrew Perry, Robin Guinn (FIN), Sandi Walters (FIN)

1.21 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 2015, staff reviewed NSUDH's current small area estimation methodology and developed small area estimation strategies for the possibly of extending the survey into Puerto Rico. Census and NSDUH staff discussed the strengths and weaknesses of these strategies. A summary of the results and findings were included in the first draft of the feasibility study report.

Staff: Robert Ashmead (x31564), Jerry Maples

1.22 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), Alisha Armas, Lauren Emanuel, Michael Hawkins, Michael Leibert, Erica Magruder, Joe Schafer, Eric Slud, Kelly Taylor

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 2015, the Integrated Research Environment (IRE) team continued to develop the IRE. The IRE will be a shared Linux computing platform that will replace the current ADRM "compute clusters" (research1, research2, and the RDC clusters.) Many common system administration tasks will be automated by reading information about "who" should have access to "what" data in the context of "which project" from the Data Management System (DMS) and CES Management System (CMS) and using that information to place people in the appropriate project and data groups. While the system was being developed, staff installed the statistical software needed for production and planned the migration. The team received "authorization to test." Testing should be underway by the end of the year.

The LAN Technology Support Office (LTSO) sponsored a group to pilot the initial rollout of R for Windows in the VDI environment. A group of volunteer R users performed load testing, and the results indicated that additional hardware is needed to support R in the VDI. Until IT obtains additional hardware, R for Windows can only be supported on Windows 7 desktops.

Staff: Chad Russell (x33215)

2. RESEARCH

2.1 GENERAL RESEARCH AND SUPPORT (Census Bureau Project 0331000)

2.2 GENERAL RESEARCH (Census Bureau Project 1871000)

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: During FY 2015, staff revised the implied ratio edit generation software GenBnds. This software implements a directed graph approach to represent edits and fields and a shortest path algorithm to generate implied edits. Staff researched updates for generating audit trail in the SAS/IML version of the computer program.

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 2015, staff supported and consulted on a wide array of projects. In the decennial area, staff implemented procedure for item and count imputation based on sequential logistic regression. Staff developed methods utilizing administrative records to identify vacant housing units and determining the census count for some of the occupied housing units. Staff completed the research on "product line imputation." Staff expanded the EM algorithm for estimating the parameters of the Mariano-Kadane model for modeling response propensities and proposed an alternative algorithm involving successive maximizations, without the need for an "E-Step." Staff revised and resubmitted a co-authored paper by to *The Annals of Applied Statistics*. Staff refereed two papers on the Fellegi-Holt model of Edit/Imputation for *JASA*. Staff provided background material to a group of individuals working on the CEDCAP project for a generalized edit/imputation methods/software for Census Bureau censuses and surveys. Staff taught the one-day short course on modeling/edit/imputation that covered background and methods related to generalized edit/imputation production software for economic censuses (Winkler 1995, 1997) and for decennial censuses and demographic surveys (Winkler 1997, 2003, 2008, 2008). Staff also covered the methods in a talk at the 2014 Fall Methodology Symposium hosted by Statistics Canada.

Staff: Yves Thibaudeau (x31706), Chandra Erdman, 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 2015, staff provided reference lists and background on research papers to Mathematical Policy Research and MITRE, who were putting together briefings/workshops for a number of government agencies including the Census Bureau. Staff provided advice to several groups related to differential privacy. Staff provided a number of examples from news sources about how knowledgeable individuals had breached privacy in several public-use (open) data situations at non-Census Bureau agencies.

One staff member continued as an Associate Editor for two computer science journals related to privacy and confidentiality. Staff provided some information to new staff at Center for Disclosure Avoidance Research (CDAR) who had previously not worked with privacy and confidentiality. In particular, staff generalized modeling/edit/imputation software to be used for generating synthetic discrete data with valid analytic properties and extremely reduced re-identification risk in contrast to differential privacy methods for which there is no generalized software.

Staff: William Winkler (x34729), William Yancey

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 finished a final draft of the chapter "Probabilistic Data Linkage" that will appear in a monograph on record linkage methods for the health sciences. Staff circulated an updated version of a large list of references to the record linkage literature.

Our *BigMatch* record linkage continues as the fastest record linkage software in the world. It is 50 times as fast as parallel software developed by computer scientists at Stanford and 500 times as fast parallel software in use at several government agencies. It is 10 times as fast as recent parallel software independently developed at the University of Leipzig and Vanderbilt University.

Staff presented in an invited paper session on record linkage at the 2014 Statistics Canada fall symposium. Staff met with six staff members at Statistics Canada who are working on new methods and production software. Staff reviewed four papers by Statistics Canada staff on record linkage and sent detailed comments about the methods and the possibility of implementing the ideas. Staff gave a talk on record linkage and modeling/edit/imputation at the Institute of Policy Research at Northwestern University.

One staff member will give the keynote address at a data integration workshop the IEEE International Conference on Data Mining and has been invited back to the Isaac Newton Institute to work on record linkage in 2016. Staff gave an Invited Roundtable on record linkage at the 2015 Joint Statistical Meetings in Seattle, WA. Staff was also the discussant in an invited paper session on record linkage at JSM.

Staff: William Winkler (x34729), William Yancey, Ned Porter

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: Staff provided information to two groups within the Census Bureau and several groups outside the Census Bureau related to methods and software for modeling/edit/imputation. Staff noted that the new methods, while seemingly better than current Census Bureau methods in preserving joint distributions, might be four to five orders of magnitude slower than our existing SPEER system (Winkler 1995, 1997) used for the last four Economic Censuses. Staff provided advice on the theoretical Integer Programming methods needed for a successful production system for economic editing.

Staff reviewed eight papers on statistical mixture models for discrete data. The review was related to how the methods related edit/imputation and to capture/recapture for record linkage. Some cruder versions of the methods were already in our DISCRETE edit/imputation system (Winkler 1997, 2008, 2010). DISCRETE is noted for extremely fast computation algorithms (two to four orders of magnitude increase over alternatives) necessary for working with sets of national files.

Staff provided comments to CSRM staff about the methods in the current modeling/edit/imputation methods and software (Winkler 2008, 2010). One staff person taught the one-day modeling/edit/imputation class.

Staff: William Winkler (x34729), William Yancey, Ned Porter, 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: Staff tested a version of the R software installed in the Census VDI environment. The test demonstrated that R can be run in multiple threads simultaneously in the VDI environment. Staff represented the group at a booth during the Affinity Group Fair.

Staff: Yves Thibaudeau (x31706), Chandra Erdman, 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 2015, staff continued research on robust methods in Small Area Estimation weakening the customary assumptions on normally distributed random area-effects, focusing on bootstrap and double-bootstrap methods of estimating Mean Squared Errors (MSEs) of small area estimation. The project led to a large simulation study of competing MSE estimation methods including bootstrap methods, which is still underway. Staff is currently running simulations to compare various factors such as underlying random effect distributions, moment structures of the data and various estimation techniques across different model frameworks. Standard errors are computed using a double bootstrap technique modified by the work of Hall to speed up the double iterative approach. Questions staff want to address include whether the double bootstrap (compared to the single bootstrap) gives better mean squared error estimates versus the additional computation complexity and how these alternative estimators compare to the standard Prasad-Rao estimator especially when the underlying assumptions are violated.

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

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 2015, staff performed an in depth analysis of the results of a factorial design simulated during FY 2014 with 1206 different configurations and 10,000 replications for each. Staff found empirical evidence from the simulation study to characterize what factors impact coverage and how each of seven intervals fare in length and coverage level across different situations. The analysis also extended to symmetry of coverage and to the distribution of design effect estimates over the simulation replications under different simulation configurations.

Staff worked on developing exhibits to effectively communicate an extensive list of conclusions based on a challenging analysis of the results. The implications to ACS were considered. Staff wrote a manuscript with the results and submitted to a journal.

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 2015, staff implemented a small area unit–level model which fits a model for the individual probability of having any disability from the SIPP data and then projecting the model onto the ACS survey data. Staff implemented exact and approximate methods for both the small area point estimate and its components for the standard error for the unit-level logistic normal model. The approximations have an added benefit in that the calculations dependent on the SIPP data are independent of the calculations dependent on the ACS data. This means that the SIPP data can be summarized without the need to access the original microdata to produce estimates and their standard errors. This could be useful when modeling sensitive outcomes where disclosure could be an issue. Staff presented some results from this research at the Joint Statistical Meetings (JSM) in August at Seattle.

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: Staff applied both a bivariate Fay-Herriot model, and, for proportions, a bivariate Binomial-Logit Normal (BLN) model to four different applications. Staff also derived theoretical results that show how the extent of standard error reduction from borrowing strength from a second survey depends on the characteristics of the underlying data. Through these theoretical calculations and real data examples, staff showed that substantial reductions in standard errors might be achieved by borrowing strength in such a manner even without covariates obtained from auxiliary information. Staff gave two invited talks on the results, one at the First Latin American ISI Satellite Meeting in Small Area Estimation in Santiago, Chile, and the other at the National Conference for Health Statistics, in Bethesda, MD.

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

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. Interviewer Observations Research
Description: An inter-agency team was formed to develop a set of interviewer observations for face-to-face surveys. The goal is to identify a set of measures that can be collected for all households in sample (both responding and non-responding households), and that predict the likelihood of both survey cooperation and key survey responses. Once such measures are identified, they may be used to manage data-collection resources and in non-response bias adjustments.

Highlights: During FY 2015, staff drafted a manuscript that illustrates the relationships between the National Health Interview Survey interviewer observations and key sample survey estimates, as well as the predictive power of the observations in response propensity models.

Staff: Chandra Erdman, Nancy A. Bates (ADRM), Peter Miller (ADRM), Rachael Walsh (CSM), Karen Ann Bagwell (CAD), James Dahlhamer (NCHS), Renee Gindi (NCHS)

B. Household Survey Design and Estimation
[See Demographic Projects]

C. 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 continued collaborating with a team in the Economic Directorate to find a statistical procedure for detecting and treating verified influential values in economic surveys to replace the current subjective procedure performed by analysts. Recent research has focused on finding an automated procedure with the expectation that any adjustments be reviewed. Previous research identified an M-estimation methodology as the most suitable choice, but the initial parameter settings for the M-estimation algorithm affect its performance. Staff continues to investigate data-driven approaches to determine the initial parameter settings for the M-estimation algorithm parameters. The goal is to select a method for a side-by-side test conducted in real time during Monthly Wholesale Trade Survey data collection.

Staff revised a research note on the Clark method of Winsorization, an alternative method for detecting and treating influential values that the team investigated before deciding to pursue the M-estimation method. Staff presented the insights gained about the performance of Clark Winsorization in detecting influential values.

Staff: Mary Mulry (x31759)

D. 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 2015, staff obtained data with estimates and associated standard errors from five countries (Brazil, Australia, South Korea, Sweden, and Israel) and developed plans for applying seven ranking tools and methods to each data set.

Staff completed and presented the analysis and slides for the 2015 60th World Statistics Congress research presentation where ranking methods were applied to data from six nations and discussed by representatives from four of these national statistical agencies.

Clarification and improvements for existing statistical methods were accomplished, and some promising methods based on the bootstrap were introduced and studied. A major objective of this research is to communicate statistical uncertainty in these rankings to wide audiences.

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

E. 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.

Highlights: During FY 2015, the exact optimal allocation method (Wright, 2012) was extended to several cases: mixed constraints on the strata sample sizes; a consideration of costs constraints; and desired precision constraints. Related methods in the literature were found and a careful study of some was begun.

Staff reviewed the literature on exact optimal sample allocation and refined two new allocations—one where the sample size is unspecified while the precision is specified, and the other where there is a budget limit.

Staff: Tommy Wright (x31702), Andrew Perry

F. Interviewer-Respondent Interactions: Gaining Cooperation

Description: Survey nonresponse rates have been increasing, leading to concerns about the accuracy of (demographic) sample survey estimates. For example, from 1990 to 2004, initial contact nonresponse rates have approximately doubled for selected household sample surveys including the Current Population Survey (CPS) (from 5.7 percent to 10.1 percent). While mailout/mailback is a relatively inexpensive data collection methodology, decreases in mailback rates to censuses and sample surveys mean increased use of methodologies that bring respondents into direct contact with Census Bureau interviewers (e.g., field representatives) using CATI (computer assisted telephone interviewing) or CAPI (computer assisted personal interviewing). CAPI can include face-to-face or telephone contact. Unsuccessful interviewer-respondent interactions can lead to increased costs due to the need for additional follow-up, and can also decrease data quality. So, unsuccessful interviewer-respondent interactions should be minimized.

This project will analyze data from 512 field representatives (interviewers) as part of an exploratory study, examining their beliefs regarding what works in gaining respondents' cooperation and investigating associations with field representatives' performance in terms of completed interview rates. We will also study associations between field representatives' beliefs and what they say they do.

Highlights: Inactive.

Staff: Tommy Wright (x31702), Tom Petkunas

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: Staff (a) finalized development of an interactive X-13ARIMA-SEATS online web application; (b) streamlined the extreme-value adjustment software; and (c) compared seasonal adjustments from SEATS and X-11 to those of a hybrid seasonal adjustment method that collects model based adjustments of moving spans of data, finding that in most cases the differences were small.

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: Staff (a) completed work on stable parametrizations of VARMA models and began extensions to constrained estimation via LASSO; (b) continued simulation and software development for multivariate count time series; (c) implemented likelihood ratio tests for Granger non-causality, as a way to exclude extraneous data from multivariate forecasting problems; (d) continued research and simulations for two tests of co-integration, one based upon fitted structural models and another based on nonparametric spectral estimates.

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

C. Time Series Model Development
Description: This work was proposed for the ASA/NSF/Census Fellowship, namely to develop a flexible integer-valued autoregressive 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 2015, staff developed an initial model and studied its statistical properties. More precisely, staff determined a form of the associated transition probability, and a probability generating function that infers the time-reversible nature of the model under consideration. Statistical computation development is underway, and further study regarding the statistical methodology is ongoing.

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. Synthetic Survey and Processing Experiments
Description: To improve operational efficiencies and reduce costs of survey processing, this project will simulate a survey, in which an artificial team of interviewers seek out an artificial set of respondents, to test alternative methods of allocating resources in the field and to test alternatives for the post-processing of the gathered survey data. When calibrated with survey paradata, the model may also serve as a test bed for new methods of missing data imputation.

Highlights: During FY 2015, staff worked with internal and external collaborators to develop a proposal for a microsimulation to study the differences between new housing records, the US Postal Service's Delivery Sequence File, and the Census Bureau's Master Address File.

Staff made a large number of minor improvements to Apophenia, a library of functions useful for statistical analysis of microsimulations. For latest version, see http://dx.doi.org/10.5281/zenodo.19255.

Staff: Ben Klemens

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: In a large sample size scenario, staff developed a rectangular tolerance region, and applied a bootstrap calibration in order to improve accuracy. If the sample size is not large, staff is pursuing a parametric bootstrap approach based on an estimated multivariate density, coupled with bootstrap calibration in order to achieve accurate coverage probabilities. The methodology is expected to provide an accurate alternative to ratio edits for identifying errors in the data. While performing ratio edits based on two correlated items, errors may go undetected if both items are erroneously recorded. However, a bivariate rectangular tolerance region can identify such errors.

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 2015, staff developed a zero-inflated COM-Poisson model and met with Decennial Statistical Studies Division (DSSD) staff to discuss variable selection in association with this construct when applied to the MAF dataset. A manuscript describing the statistical methodology associated with this work has been revised and is under review with a statistics journal. Results discussed in this work include the ability of this zero-inflated COM-Poisson regression to account for excess zeroes and address data dispersion, and its ability to generalize zero-inflated Poisson and zero-inflated geometric regressions, respectively (as well as logistic regression).

Staff: Kimberly Sellers (x39808), Andrew Raim

D. Modeling the Causal Effects of Field Representatives' Actions and Strategies

Description: Field Representatives (FRs) apply different strategies for managing their monthly workloads. For example, some FRs may place high priority on contacting households that are perceived as likely to respond, putting aside the more difficult cases until later in the month. With large volumes of information flowing from paradata systems, we are better able to model FR data collection behavior. However, to understand the causal effects of these behaviors on outcomes of interest (response rates, measures of data quality and measures of cost), we need to adjust for confounding characteristics of FRs and their caseloads. In this project, we are developing techniques for causal inference from observational (non-experimental) data on FR characteristics, behaviors and performance measures.

Highlights: During FY 2015, staff developed new methods for causal inference with a continuous treatment variable using a framework of estimating functions. Staff compared them to existing methods for continuous treatments and evaluated their performance in a simulated population where the true causal dose-response function is linear. These findings were summarized in a draft research report that is being prepared for submission to a peer reviewed journal.

Staff evaluated the newly developed estimators by applying them to two non-survey data examples: a study of the effect of winning the lottery on earnings that has been repeatedly analyzed in the literature, and a longitudinal study linking declining physical activity in adolescent females to physical activity and body mass index in young adulthood. Staff began the process of bundling our software routines as an R package. Staff have had several meetings with staff from the newly formed Office of Survey and Census Analytics who are engaged in exploratory causal analyses of survey process data, and made a presentation on causal inference at the monthly meeting of Designing Effective Experiments Project (DEEP).

Staff: Joe Schafer (x31823), Doug Galagate, Robert Ashmead

E. 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 2015, staff studied the statistical properties of the bivariate form of the Conway-Maxwell-Poisson distribution. While this distribution is able to account for data over- or under-dispersion, the model can only capture data dispersion of the same form for the two random variables. The model can be used as an exploratory analysis tool for model selection because it captures the bivariate Poisson, bivariate Bernoulli, and bivariate geometric distributions as special cases; but further may possibly be used to aid with statistical disclosure limitation for sensitive bivariate count data containing significant data dispersion. A research manuscript has been accepted for publication in the *Journal of Multivariate Analysis.* Staff applied the distribution to an administrative records project in an effort to describe the distributional form of bivariate information stemming from census data and administrative records.

Staff: Kimberly Sellers (x39808), Darcy Steeg Morris

F. 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 2015, staff submitted a manuscript for publication and continue to revise the manuscript and supporting codes for the journal in question.

Staff: Kimberly Sellers (x39808), Darcy Steeg 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 computationally intensive 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 by 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: Staff developed new likelihood based procedures for drawing inference based on multiply imputed partially synthetic data under the univariate normal model. Staff considered synthetic data generated via two methods: plug-in sampling and posterior predictive sampling. Under plug-in sampling, synthetic data are generated from the assumed parametric model with unknown parameters set equal to the observed values of their point estimators based on the actually observed data. Under posterior predictive sampling, data are generated from an appropriate posterior predictive distribution in a Bayesian setting. Simulation studies were conducted to show that these new methods performed as our theory predicted. These results are described in the manuscript entitled "Likelihood-Based Inference for Singly and Multiply Imputed Synthetic Data under a Normal Model." This manuscript was revised and submitted for publication, and subsequently accepted for publication in *Statistics and Probability Letters*.

A manuscript describing the work on likelihood based inference for singly imputed data under posterior predictive sampling was written, submitted, and accepted for publication in *Sankhya B: The Indian Journal of Statistics*. This manuscript is entitled "Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models."

Staff continued work on the development of new and exact methods for drawing parametric inference based on singly imputed partially synthetic data generated via plug-in sampling. Staff conducted a series of simulation studies under multivariate normal and multiple linear regression models to demonstrate that the developed methodology did in fact perform as our theory

predicted. Staff also conducted simulation studies to compare the accuracy of inference of singly imputed partially synthetic data analyzed using our methods with the accuracy of inference of multiply imputed partially synthetic data analyzed using well established combining formulas. As an example, staff applied our methodology for singly imputed partially synthetic data, as well as established methods for multiply imputed partially synthetic data, to public use data from the Current Population Survey (CPS). Finally, staff performed a disclosure risk evaluation of singly and multiply imputed partially synthetic data in the context of the CPS example. Staff found that our methods gave valid statistical inferences, but multiply imputed synthetic data yield more efficient inference than singly imputed partially synthetic data. Staff alternatively found that single imputation can enhance the level of privacy protection in comparison with multiple imputation. Staff then extended the methodology by: (1) considering a scenario where only a portion of the data are sensitive, as determined by a vector of covariates. Under this setup, staff developed an exact data analysis method when only these sensitive values are synthesized (using single imputation), while the non-sensitive values are left unperturbed, (2) developing an exact data analysis method under the scenario when both the response variable, as well as the covariate vector are sensitive. Here, staff assume that the released data consist of a singly imputed synthetic version of both the response variable and covariates and we develop exact methods to analyze these released data, assuming the original values follow a multivariate normal distribution. Staff also began to study how methodology performs under certain types of model misspecification. Staff is documenting this work in a manuscript entitled, "Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models," which is in progress.

Staff developed new methodology for drawing likelihood based inference based on both singly and multiply imputed synthetic data (generated via plug-in sampling) under a multivariate linear regression model. In this setting, the parameter of interest is the unknown matrix of regression coefficients. Staff conducted simulation studies to demonstrate that the methodology performed as the theory predicted, and as an illustration, applied the methods to public use data from the Current Population Survey. Staff also conducted a disclosure risk evaluation of singly versus multiply imputed synthetic data in this specific setting of a multivariate linear regression model. This work is documented in the manuscript, "Inference for Multivariate Regression Model based on Synthetic Data generated using Plug-in Sampling," which is in preparation.

Staff initiated a study on the use of bootstrap in the context of statistical disclosure control. Specifically, we used simulation to study the quality of bootstrap inferences based on noise-multiplied data and singly imputed partially synthetic data generated via plug-in sampling. Preliminary results under some simple parametric models suggest that bootstrap can yield valid inference in these cases. Staff also investigated bootstrap in the case where a singly imputed synthetic dataset is generated using Classification and Regression Trees.

Staff began working on a new procedure for using noise multiplication and noise addition to protect nonnegative data whose distribution is described by a point mass at zero mixed with a continuous distribution having positive support (such distributions can often occur in practice). A Bayesian approach to the problem has been outlined, and staff is investigating Markov chain Monte Carlo based methods for conducting inference.

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 present a seminar based on their research and engage in collaborative research with Census Bureau researchers and staff.

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

Highlights: During FY 2015, CSRM staff worked with staff across the Census Bureau and hosted the visits of thirty *2015 SUMMER AT CENSUS* scholars. See the listing under Section 5 for this year's program.

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: Alisha Armas, 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. (In Press). "Groupwise Sufficient Dimension Reduction with Principal Fitted Components," *Computational Statistics.*

Adragni, K. P., and Raim, A. M. (2014). "Idr: An R Software Package for Likelihood-Based Sufficient Dimension Reduction," *Journal of Statistical Software, 61(3).*

Dong, X. and Mathew, T. (2015). "Central Tolerance Regions and Reference Regions for Multivariate Normal Populations," *Journal of Multivariate Analysis, 134,* 50-60.

Franco, C. and Bell, W.R. (In Press). "Borrowing Information over Time in Binomial/Logit Normal Models for Small Area Estimation of Poverty," *Statistics in Transition* and *Survey Methodology*.

Janicki, R. and McElroy, T. (In Press). "Hermite Expansion and Estimation of Monotonic Transformations of Gaussian Data," *Journal of Nonparametric Statistics*.

Klein, M., and Sinha, B. (2015). "Likelihood-Based Finite Sample Inference for Synthetic Data Based on Exponential Model," *Thailand Statistician: Journal of the Thai Statistical Association, 13:* 33-47.

Klein, M. and Sinha, B. (In Press). "Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models," *Sankhya B: The Indian Journal of Statistics.*

Klein, M. and Sinha, B. (2015). "Likelihood-Based Inference for Singly and Multiply Imputed Synthetic Data under a Normal Model," *Statistics and Probability Letters, 105,* 168-175.

Livsey, J., Kechagias, S., Lund, R., and Pipiras, V. (In Press). "A Long-Memory Multivariate Model for Count Time Series with Flexible Autocovariance Structure," *Annals of Applied Statistics.*

Malinovsky, Y., Albert, P., and Roy, A. (2015). "A Note on the Evaluation of Group Testing Algorithms in the Presence of Misclassification," *Biometrics.*

Mbodj, M. and Mathew, T. (2015). "Approximate Ellipsoidal Tolerance Regions for Multivariate Normal Populations," *Statistics and Probability Letters, 97,* 41-45.

McElroy, T. (2015). "When are Direct Multi-Step and Iterative Forecasts Identical?" *Journal of Forecasting, 34*: 315-336.

McElroy, T. (2015). "Multivariate Seasonal Adjustment, Economic Identities, and Seasonal Taxonomy," Forthcoming, *Journal of Business and Economics Statistics*.

McElroy, T. and McCracken, M. (2015). "Multi-Step Ahead Forecasting of Vector Time Series," Published online, *Econometrics Reviews.*

McElroy, T. and Monsell, B. (2015). "Model Estimation, Prediction, and Signal Extraction for Nonstationary Stock and Flow Time Series Observed at Mixed Frequencies," *Journal of the American Statistical Association (Theory and Methods), 110:* 1284-1303.

McElroy, T. and Nagaraja, C. (2015). "Tail Index Estimation with a Fixed Tuning Parameter Fraction," Published online, *Journal of Statistical Planning and Inference.*

McElroy, T. and Trimbur, T. (2015). "Signal Extraction for Nonstationary Multivariate Time Series with Illustrations for Trend Ination," *Journal of Time Series Analysis, 36*: 209-227.

Nagaraja, C. and McElroy, T. (2015). "On the Interpretation of Multi-Year Estimates of the American Community Survey as Period Estimates," Published online, *Journal of the International Association of Official Statistics*.

Raim, A., Neerchal, N., and Morel, J. (In Press). "An Approximation to the Information Matrix of Exponential Family Finite Mixtures," *Annals of the Institute of Statistical Mathematics*.

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

Young, D. and Mathew, T. (2015). "Ratio Edits Based on Statistical Tolerance Intervals." *Journal of Official Statistics, 31(1):* 77-100.

3.2 BOOKS/BOOK CHAPTERS

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

3.3 PROCEEDINGS PAPERS

Joint Statistical Meetings, American Statistical Association, Boston, MA, August 2-7, 2014.
2014 Proceedings of the American Statistical Association

- Yang Cheng, Adrijo Chakrabarty, and Guari Datta, "Hierarchical Bayesian Methods for Combining Surveys", 4099-4111.
- Jiashen You, Guari Datta, and Jerry Maples, "Modeling Disability in Small Areas: An Area-Level Approach of Combining Two Surveys", 3770-3784.
- Carolina Franco, Roderick Little, Thomas Louis, and Eric Slud, "Coverage Properties of Confidence Intervals for Proportions in Complex Sample Surveys", 1799-1813.
- Martin Klein, Joanna Fane Lineback, and Joseph Schafer, "Evaluating Imputation Techniques in the Monthly Wholesale Trade Survey", 1814-1827.
- James Livsey, Tucker McElroy, and Anindya Roy, "Guassianity versus Correlation in Time Series Residual Diagnostics", 3334-3343.
- Tucker McElroy and Scott Holan, "Fast Estimation of Time Series with Multiple Long-Range Persistencies", 13-27.
- Brian Monsell, "The Effect of Forecasting on X-11 Adjustment Filters", 1368-1381.
- Darcy Steeg Morris, "A Comparison of Methodologies for Classification of Administrative Records Quality for Census Enumeration", 1729-1743.
- Mary Mulry, Elizabeth Nichols, and Jennifer Hunter Childs, "Study of Error in Survey Reports of Move Month Using the U.S. Postal Service Change of Address Records", 3109-3123.
- Andrew Raim, Nagaraj Neerchal, and Jorge Morel, "Large Cluster Approximation to the Finite Mixture Information Matrix with an Application to Meta-Analysis", 4025-4037.
- Eric Slud, "Modeling Frame Deficiencies for Improved Calibrations", 1774-1786.
- Brian Shaffer, Yang Cheng, and Eric Slud, "Single-Stage Generalized Raking Application in the American Housing Survey", 3708-3719.
- William Winkler, "Quality and Analysis of Sets of National Files", 1432-1442.

3.4 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY RESEARCH REPORTS

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

RR (Statistics #2014-10): Jerry J. Maples and Matthew Brault. "Improving Small Area Disability: Combining the American Community Survey with the Survey of Income and Program Participation," November 12, 2014.

RR (Statistics #2014-11): Jiashen You, Guari S. Datta, and Jerry J. Maples. "Modeling Disability in Small Areas: An Area-Level Approach of Combining Two Surveys," November 12, 2014.

RR (Statistics #2014-12): Tommy Wright, Martin Klein, and Jerzy Wieczorek. "Ranking Populations Based on Sample Survey Data," November 14, 2014.

RR (Statistics #2015-01): Tucker S. McElroy. "Multivariate Seasonal Adjustment, Economic Identities, and Seasonal Taxonomy," February 26, 2015.

RR (Statistics #2015-02): David F. Findley, Demetra P. Lytras, and Augustin Maravall. "Illuminating Model-Based Seasonal Adjustment with the First Order Seasonal Autoregressive and Airline Models," July 10, 2015.

RR (Statistics #2015-03): William R. Bell. "Unit Root Properties of Seasonal Adjustment and Related Filters: Special Cases," June 30, 2015.

3.5 OTHER REPORTS

Raim, A.M., Neerchal, N.K., and Morel, J.G. (2015). "Modeling Overdispersion in R," *Technical Report HPCF-2015-1, UMBC High Performance Computing Facility*, University of Maryland, Baltimore County. 2015.

Thompson, K.J., Liu, X., Bechtel, L., Diamond K., Davie, W. Dorsett, F., Ellis, Y., Garcia, M., Kern, J., Knutson, J., Martin, J., Morris, D.S., Schuyler J., Strubble, R., Tolliver, K., and Ward, J. (2014). "Recommendation for Product Line Imputation for 2017 Economic Census," *Economic Statistical Methods Division Research Report*. US Census Bureau. Washington DC.

Slud, E. (2015). "Impact of Mode-based Imputation on ACS Estimates," *American Community Survey Research and Evaluation Memorandum*, #ACIS-RER-O7.

4. TALKS AND PRESENTATIONS

2014 International Total Survey Error Workshop, Washington, D.C., October 1-3, 2014.

- Mulry, Mary, "Are Proxy Responses Better than Administrative Records."

Statistics Department Colloquium, University of Virginia, Charlottesville, Virginia, October 10, 2014.

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

Statistics Canada 2014 International Symposium on Methodology in Statistics, Hull, Quebec, Canada, October 29-31, 2014.

- Winkler, William, "Quality and Analysis of National Files."

Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia, November 6, 2014.

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

Statistical Seminar, Centro de Investigación Cooperativa, Universidad Miguel Hernández de Elche, Elche, Spain. November 7, 2014.

- Franco, Carolina (based on joint work with William R. Bell), "On Borrowing Information over Time in Small Area Estimation with Application to Small Area Income and Poverty Estimates (SAIPE)."

2014 International Statistics Conference, Colombo, Sri Lanka, December 28-30, 2014.

- Livsey, James, "Periodicities and a Multivariate Model for Count Time Series with Flexible Autocovariance Structure."

Institute for Integrating Statistics in Decision Sciences, Department of Decision Sciences, The George Washington University, Washington, D.C., January 2015.

- Sellers, Kimberly, "Don't Count on Poisson! Introducing the Conway-Maxwell-Poisson Distribution to Model Count Data."

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, Maryland, February 6, 2015.

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

Northwestern University, Institute for Policy Research and Department of Statistics, Evanston, Illinois, February 11, 2015.

- Winkler, William, "Quality and Statistical Analysis of Sets of National Files."

Business Week, University of Texas at Arlington, Arlington, Texas, March 16 – 20, 2015.

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

Statistical Consulting Seminar, Department of Statistical Sciences, Southern Methodist University, Dallas, Texas, April 6, 2015.

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

National Heart, Lung, and Blood Institute/National Institutes of Health, Office of Biostatistics Research, Bethesda, Maryland, April 14, 2015

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

9th Annual Probability and Statistics Day at UMBC, Special Alumni Session, Baltimore, Maryland, April 18, 2015.

- Raim, Andrew, "Mixture Link Models for Binomial Data with Overdispersion."

X-STEM Symposium, USA Science & Engineering Festival, Washington, D.C., April 28, 2015.

- Sellers, Kimberly, "CSI: Count on Statistics for Investigation."

George Mason University, Fairfax, VA, May 1, 2015.

- Livsey, James, "Renewal Methods for Integer-valued Time Series."

Bureau of Labor Statistics, Washington, D.C., May 14, 2015.

- Livsey, James and Monsell, Brian, "Update on X-13ARIMA-SEATS Project."

Workshop on Sampling (Sponsored by ICDDR-B; USAID—Bangladesh; U.S. Census Bureau), International Center for Diarrheal Disease Research, Dhaka, Bangladesh, May 17-21, 2015.

- Wright, Tommy, Conducted entire workshop presenting the main components of survey sampling with a focus on probability sampling techniques.

2015 International Symposium on Forecasting, Riverside, California, June 22-24, 2015.

- McElroy, Tucker, "The Application of Hysteric Decompositions to Signal Extraction."

Fourth Italian Conference on Survey Methodology, Invited Session on Poverty Mapping, Rome Italy, June 25, 2015.

- Franco, Carolina (based on joint work with William R. Bell), "Estimating Poverty Rates of School-Aged Children for US Counties through Cross-Sectional and Temporal Small Area Estimation Models."

38th Conference on Stochastic Processes and Applications, University of Oxford, Oxford, UK, July 13-17, 2015.

- Sellers, Kimberly, "A Flexible Stochastic Process for Count Data."

60th World Statistics Congress of the International Statistical Institute, Rio de Janeiro, Brazil, July 26-31, 2015.

- Erdman, Chandra, "Predicting Response Mode During Data Collection in the National Survey of College Graduates."
- Wright, Tommy and Klein, Martin, "Towards Expressing Statistical Uncertainty in Rankings Based on Sample Survey Data from Six National Statistical Agencies."

First Latin American ISI Satellite Meeting on Small Area Estimation, Santiago, Chile, August 3-5, 2015.

- Franco, Carolina (based on joint work with William R. Bell), "Continuing Estimates from Related Surveys via Bivariate Area-Level Models without Covariates."

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

- Robert Ashmead and Bo Lu, "Health Care Policy Evaluation Using Propensity Score Matching: A Study of Care Consistent with a Patient-Centered Medical Home Using a Large Population Survey."
- Soutir Bandyopadhyay, Tucker McElroy, and Doug Nychka, "Spatial Modeling of the American Community Survey."
- William R. Bell and Carolina Franco, "On Borrowing Information Over Time in Small-Area Estimation."
- Adrijo Chakraborty, Gauri Datta, and Abhyuday Mandal, "Robust Bayesian Small-Area Estimation for Area-Level Data."
- James Dahlhamer, Renee Gindi, Chris Moriarity, and Chandra Erdman, "Assessing the Utility of Interviewer Observations for Nonresponse Adjustments in the National Health Interview Survey."
- Gauri Datta, Aurore Delaigle, Peter Hall, and Lily Wang, "Prediction Intervals of Small-Area Means Under Semiparametric Measurement Error Models."
- Chandra Erdman, "Examining the Predictive Power of Response Propensity Models in Varied Survey Designs."
- Maria Garcia, Darcy Steeg Morris, and L. Kaili Diamond, "Implementation of Ratio Imputation and Sequential Regression Multivariate Imputation on Economic Census Products."
- Ryan Janicki, "Estimation of the Difference of Small-Area Parameters from Different Time Periods."
- Martin Klein, "On a Comparison of Singly and Multiply Imputed Partially Synthetic Data Under Plug-In Sampling."
- Joanna Lineback, Martin Klein, and Joseph Schafer, "Evaluating Estimation Techniques in the Monthly Wholesale Trade Survey."
- James Livsey, "Interactive X-13-ARIMA-SEATS."

- Jerry Maples and Amy Steinweg, "Improving Small-Area Estimates of Disability: A Model-Based Approach to Combining the American Community Survey with the Survey of Income and Program Participation."
- Brian Monsell, David Findley, and Osbert Pang, "A Moving Window Approach to Model-Based Seasonal Adjustment."
- Darcy Steeg Morris, "Administrative Record Research to Reduce Contacts in the 2020 Census."
- Mary Mulry, Andrew Keller, and Tyler Fox, "Are Proxy Responses Better Than Administrative Records."
- Osbert Pang and Brian Monsell, "Examining Diagnostics for Trading-Day Effects from X-13-ARIMA-SEATS."
- Andrew Raim, Nagaraj Neerchal, and Jorge Morel, "Mixture Link Models for Binomial Data with Overdispersion."
- Anindya Roy and Tucker McElroy, "Moment-Based Estimation of VARMA Parameters."
- Eric Slud, "Weighted Estimating Equations Based on Response Propensities in Terms of Covariates That Are Observed Only for Responders."
- Bill Winkler (Discussant for Session), "Bayesian Approaches to Record Linkage."
- Bill Winkler, "Record Linkage: Introductory Overview" (Roundtable Discussion).
- Tommy Wright, "A Simple and General Algorithm for Exact Optimal Sample Allocation That Is More Efficient Than Neyman Allocation."
- Jiashen You and Gauri Datta, "Mixture Model and EM Algorithm in Small-Area Estimation."

National Conference on Health Statistics, Bethesda, Maryland, August 24-26, 2015.

- Franco, Carolina (based on joint work with William R. Bell), "Using ACS Estimates to Improve Estimates from Other US Surveys."

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, Maryland, September 4, 2015.

- Franco, Carolina (based on joint work with William R. Bell), "Borrowing Information over Time in Binomial/Logit Normal Models for Small Area Estimation."

2015 International Conference on Total Survey Error, Baltimore, Maryland, September 20-22, 2015.

- Ashmead, Robert, "Examining Mode Options for the Commodity Flow Survey."
- Mulry, Mary and Keller, Andrew, "Comparing the Quality of 2010 Census Proxy Responses with Administrative Records."
- Nichols, Elizabeth M., Mulry, Mary, and Childs, Jennifer H., "Using Administrative Records Data at the U.S. Census Bureau: Lessons Learned from Two Research Projects Evaluating Survey Data."
- Raim, Andrew, "Selection of Predictors to Model Coverage Errors in the Master Address File."

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

Andrew E. Hong, University of Pennsylvania, "Anomaly Detection in Earth Satellite Imaging," October 14, 2014.

Ross Hilton, Georgia Institute of Technology, "Profiling Utilization from Large, Highly-Sensitive Medical Claims Data," November 18, 2014.

Douglas VanDerwerken, Duke University, "A Hodgepodge of Novel Bayesian Methods," December 17, 2014.

Stephanie Eckman (Former U.S. Census Bureau Dissertation Fellow), Institut für Arbeitsmarkt- und Berufsforschung (IAB), (Institute for Employment Research), "Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias?" January 6, 2015.

Michael Tzen, DSMD, U.S. Census Bureau, "Spatial & Tree Methods on Big Data Master Address File," January 22, 2015.

Xingyou Zhang, Centers for Disease Control and Prevention (CDC), "Multilevel Regression and Poststratification (MRP) for Small Area Estimation of Health Outcomes Using Geocoded National Health Surveys for Public Health Practice," February 12, 2015.

Martin Klein, CSRM, U.S. Census Bureau, "Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-normal Regression Samples," February 19, 2015.

Chandra Erdman, CSRM, U.S. Census Bureau, "Commonalities in Survey Response Propensity Models," February 24, 2015.

Stephanie Chan-Yang, University of California, Davis, "A Maximum Entropy Approach to Joint Modeling Primate Multiple Behavior Social Networks and a Classification of Frog Species from Audio Calls," April 9, 2015.

Kimberly Sellers, Georgetown University (ASF/NSF/Census Research Fellow), "Don't Count on Poisson: Accounting for Dispersion in Count Data Modeling," April 21, 2015.

Elias Al-Najjar, University of Maryland, Baltimore County, "Hierarchical Principal Fitted Components," April 23, 2015.

Ben Klemens, U.S. Census Bureau, "Controlling Revisions on Linux," May 7, 2015.

Emanuel Ben-David, Columbia University, "Estimations in Graphical Models," May 12, 2015.

Huilin Li, New York University, *SUMMER AT CENSUS*, "An Adjusted Maximum Likelihood Method for Solving Small Area Estimation Problems," May 19, 2015.

Thomas Trimbur, George Washington University (Adjunct), "General Multivariate Low—Pass and Band-Pass Filters for Extracting Trends and Cycles in Multiple Time Series with Applications to U.S. Macroeconomic Data," May 28, 2015.

Hiroma Ishizawa, George Washington University, *SUMMER AT CENSUS*, "Dynamics of Spanish-Language Neighborhoods in Chicago and Atlanta: 1990-2000," June 2, 2015.

Christoph Sax, Consultant, "R Shiny and Seasonal: R Interface to X-13," June 3, 2015.

Thomas Mathew, CSRM, U.S. Census Bureau & University of Maryland, Baltimore County, "Series of Lectures in Experimental Design," June 4, 11, 18, 25, and July 9, 15, 16, 30, 2015.

Matthew Hall, Cornell University, *SUMMER AT CENSUS*, "Racial Separation at Home and Work: Exploring Patterns of Trends in Segregation in Residential and Workplace Contexts," June 8, 2015.

Ian Schmutte, University of Georgia, *SUMMER AT CENSUS*, "Estimating Compensating Wage Differentials with Endogenous Job Mobility," June 8, 2015.

James R. Lewis, IBM Corporation, *SUMMER AT CENSUS*, "Survey of Questionnaires Developed to Assess Satisfaction with Computers and Web Applications," June 9, 2015.

Amanda Kowalski, Yale University, *SUMMER AT CENSUS*, "Medicaid as an Investment in Children: What is the Long-Term Impact on Tax Receipts?" June 9, 2015.

James R. Lewis, IBM Corporation, *SUMMER AT CENSUS*, "How to Develop a Standardized Satisfaction Questionnaire," June 9, 2015.

Omer Ozturk, The Ohio State University, *SUMMER AT CENSUS*, "Estimation of Mean and Total of a Finite Population Using Inclusion Probabilities Give the Population Ranks," June 15, 2015.

Dongchu Sun, University of Missouri-Columbia, *SUMMER AT CENSUS*, "Hierarchical Models for Small Area Estimation with Applications in Sampling Survey," June 16, 2015.

Dongchu Sun, University of Missouri-Columbia, *SUMMER AT CENSUS*, "Bayesian Analysis of Multivariate Correlated Curves," June 16, 2015.

Jae Kwang Kim, Iowa State University, *SUMMER AT CENSUS*, "Fractional Hot Deck Imputation for Multivariate Missing Data," June 23, 2015.

Jae Kwang Kim, Iowa State University, *SUMMER AT CENSUS*, "Improving State-Level Estimation Using Area Level Models Incorporating Several Sources of Information," June 24, 2015.

Jorge Duany, Florida International University, *SUMMER AT CENSUS*, "Racial Identity in the Hispanic Caribbean and the United States," June 30, 2015.

Bikas Sinha, Indian Statistical Institute (Retired Professor), *SUMMER AT CENSUS*, "Eliciting Information on Sensitive Features: Block Total Response Technique and Related Inference," July 6, 2015.

Eitan Greenshtein, Central Bureau of Statistics – Israel, *SUMMER AT CENSUS*, "Deconvolution, Empirical Bayes, and Treatment of Non-response," July 13, 2015.

Soumendra Lahiri, North Carolina State University, *SUMMER AT CENSUS*, "Spatial Statistics, Empirical Likelihood and Survey Data," July 20, 2015.

Dan Nordman, Iowa State University, *SUMMER AT CENSUS*, "An Empirical Likelihood Method for Geo-referenced Spatial Data," July 21, 2015.

Soutir Bandyopadhyay, Lehigh University, *SUMMER AT CENSUS*, "Wendland Meets Markov: Kriging for LARGE Spatial Data," July 21, 2015.

Kyle Herkenhoff, University of Minnesota, *SUMMER AT CENSUS*, "How Credit Constraints Impact Job Finding Rates, Sorting & Aggregate Output," July 27, 2015.

Shigehiro Oishi, University of Virginia, *SUMMER AT CENSUS*, "National Statistics on Residential Mobility: How Researchers are Using Them to Understand Societal Trends," July 27, 2015.

Ming Ji, University of South Florida, *SUMMER AT CENSUS*, "Pattern Mixture Weighted GEE Models for Survey Nonresponse Using SAS PROC GENMOD," July 28, 2015.

Carolyn Liebler, University of Minnesota, *SUMMER AT CENSUS*, "On the Boundaries of Race: Identification of Mixed-Heritage Children in the U.S., 1970 to 2010," July 28, 2015.

Daniel Goldberg, Texas A&M University, *SUMMER AT CENSUS*, "Fitness for Geo-Use: Interpreting Quality in Geocoding Systems," August 3, 2015.

John W. Emerson, Yale University, *SUMMER AT CENSUS*, "An Intensive R Workshop," August 3-7, 2015.

Otavio Bartalotti, Iowa State University, *SUMMER AT CENSUS*, "Estimation and Inference in Regression Discontinuity Designs with Clustered Sampling," August 4, 2015.

John W. Emerson, Yale University, *SUMMER AT CENSUS*, "Advances in Data Visualization with R," August 5, 2015.

Wei Lin, Capital University of Economics and Business (CUEB) – Beijing, China, Texas A&M University, *SUMMER AT CENSUS*, "Functional-PCA Based Seasonal Adjustment Time Series Models with Mixed Frequencies Data," August 11, 2015.

Jennifer Dykema, University of Wisconsin-Madison, *SUMMER AT CENSUS*, "The Role of Question Characteristics in Designing and Evaluating Survey Questions," August 18, 2015.

Randall Akee, University of California, Los Angeles, *SUMMER AT CENSUS*, "Errors in Self-Reporting Earnings: The Role of Previous Earnings Volatility and Individual Characteristics," August 19, 2015.

Benjamin Pugsley, Federal Reserve Bank of New York, *SUMMER AT CENSUS*, "Understanding the 30-year Decline in Business Formation: A General Equilibrium Approach," August 25, 2015.

Aliya Saperstein, Stanford University, *SUMMER AT CENSUS*, "Trends in the Measurement of Race in U.S. National Surveys," September 1, 2015.

Yongjin Jin, Renmin University – Beijing, China, *SUMMER AT CENSUS*, "Sampling Design for Chinese General Social Survey," September 8, 2015.

Andrew Penner, University of California, Irvine, *SUMMER AT CENSUS*, "The Organization of Gender Inequality," September 15, 2015.

Thurston Domina, University of California, Irvine, *SUMMER AT CENSUS*, "'Membership Has Its Privileges': Status Incentives and Categorical Inequality in Education," September 15, 2015.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

2015 Health Policy Statistics Section (ASA) Student Paper Award

- **Robert Ashmead** – "For excellence in a methodological innovation or a creative application of statistical analysis to an issue in health policy."

6.2 SIGNIFICANT SERVICE TO PROFESSION

Robert Ashmead
- Judge, University of Maryland, Baltimore County Probability and Statistics Day
- Refereed a paper for *Journal of Survey Statistics and Methodology*

Carolina Franco
- Organizer, Invited Session, "Interesting Applications of Small Area Estimation." First Latin American ISI Satellite Meeting on Small Area Estimation, August 2015 in Santiago, Chile
- Member, Award Selection Committee, NSF Student Travel Award to attend the 2015 ISI Satellite Meeting on Small Area Estimation

Maria Garcia
- Member, Organizing Committee, 2015 Federal Committee on Statistical Methodology Research Conference

Ryan Janicki
- Refereed a paper for *The American Statistician*
- Reviewer, Grant Proposal, National Science Foundation

Patrick Joyce
- Refereed a paper for *Survey Methodology*

Martin Klein
- Refereed papers for *Statistics and Probability Letters, Computational Statistics and Data Analysis,* and *Statistical Papers*
- Member, Two Ph.D. Dissertation in Statistics Committees, University of Maryland, Baltimore County

James Livsey
- Refereed papers for *The American Statistician, International Journal of Forecasting,* and *Communications in Statistics*

Jerry Maples
- Refereed a paper for *Journal of Official Statistics*

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 papers for *Journal of Biopharmaceutical Statistics, Technometrics, Journal of Statistical Computation and Simulation, Statistics and Probability Letters, Statistics in Medicine*

Tucker McElroy

- Refereed papers for *The American Statistician, Statistical Inference for Stochastic Processes, Annals of Statistics, Journal of Time Series Analysis, Journal of Official Statistics, Communications in Statistics*, and a monograph under contract at Springer
- Co-Advisor, Two Ph.D. Students, one at University of Maryland, Baltimore County and one at George Mason University

Darcy Steeg Morris

- Reviewed papers for *The American Statistician* and *Biometrics*

Mary Mulry

- Associate Editor, *Journal of Official Statistics*
- Invited Session Organizer, Fifth International Conference on Establishment Surveys (ICES-V)

Andrew Raim

- Project Client Participant, NSF Research Experience for Undergraduates (REU), University of Maryland, Baltimore County
- Refereed a paper for *Journal of Statistical Computation and Simulation*
- Member, Ph.D. Dissertation in Statistics Committee, University of Maryland, Baltimore County

Joe Schafer

- External Reviewer, Tenure and Promotion for Faculty, The Ohio State University
- Member, Ph.D. Dissertation Committee, University of Maryland, College Park

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
- Reviewed manuscripts 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*

Eric Slud

- Associate Editor, *Journal of Survey Statistics and Methodology*
- Associate Editor, *Biometrika*
- Associate Editor, *Lifetime Data Analysis*
- Member and Chair, Hansen Lecture Committee of the Washington Statistical Society

William Winkler

- Refereed papers for *Journal of the American Statistical Association, Journal of Survey Statistics and Methodology, ACM KDD Population Informatics,* and *IEEE ICDM Data Integration and Applications*
- Associate Editor, *Journal of Privacy and Confidentiality*
- Associate Editor, *Transactions on Data Privacy*
- Member, Program Committee for *Statistical Data Protection 2015*
- Member, Program Committee for *Statistical Data Protection 2016*
- Member, Program Committee for *IEEE 2015 ICDM Data Integration and Applications Workshop*
- Member, Ph.D. in Statistics Review Committee, Carnegie-Mellon University
- External Reviewer for Promotion to Professor, Machine Learning Department in the School of Computer Science at Carnegie-Melon University and at the School of Computer Science at Australia National University

Tommy Wright

- Associate Editor, *The American Statistician*
- Member, AAAS Review Panel for 2015 Science and Technology "Big Data" Fellows

- Member, Waksberg Award Committee, Survey Methodology
- Member, Board of Trustees, National Institute of Statistical Sciences
- Member, Committee on Community-based Flood Insurance Options, National Research Council
- Member, 2015 Jan Tinberger Award Committee, International Statistical Institute
- Organizer, Invited Paper Session—"Rankings by Statistical Agencies Based on Sample Survey Data," 60th World Statistics Congress/International Statistical Institute (ISI), Brazil (2015)

6.3 PERSONNEL NOTES

Aaron Gilary accepted a position in the Demographic Statistical Methods Division.

Jiashen You accepted a position in the Center for Disclosure Avoidance Research.

Lauren Emanuel joined our center as Scientific Editor.

Andrew Perry (sophomore mathematics major at University of Maryland, College Park) joined our center as an intern.

Xiaoyun Lu joined our Machine Learning and Computational Statistics Research Group.

Ben Klemens accepted a position with the U.S. Treasury Department.

Marissa Gargano joined our Experimentation and Modeling Research Group as an intern and employee before accepting a position in private industry.

Thomas Trimbur (adjunct faculty at local universities) joined our Time Series Research Group on a Schedule A appointment.

Joe Schafer accepted a position in the Research and Methodology Directorate Office.

Chandra Erdman accepted a position with a non-profit.

Emanuel Ben-David joined our Machine Learning & Computational Statistics Research Group.

Bill Yancey retired after 24 years of federal service.

Kimberly Sellers (faculty at Georgetown University) completed an ASA/NSF/Census Research Fellowship with us and joined our Experimentation and Modeling Research Group on a Schedule A appointment.

APPENDIX A **Center for Statistical Research and Methodology FY 2015**
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		
6410A00-D03	Coverage Improvement Design		
6410A00-D22	Non Response Follow-up Design and Operations		
	1. Coverage Measurement Research	Jerry Maples	Pat Cantwell
	2. Using 2010 Census Coverage Measurement Data to Compare Nonresponse Follow-up Proxy Responses with Administrative Records	Mary Mulry	Tom Mule
	3. Supplementing and Supporting Non-Response with Administrative Records	Michael Ikeda	Tom Mule
	4. Identifying "Good" Administrative Records for 2020 Census NRFU Curtailment Targeting	Darcy Steeg Morris	Tom Mule
	5. Census Cost-Quality Tradeoff Assessment	Michael Ikeda	Andreana Able
6310A00-C02	Address List Quality Measurement		
	6. Master Address File (MAF) Error Model and Quality Assessment	Andrew Raim	Laura Ferreira
	7. Development of Block Data Tracking Database	Tom Petkunas	Michael Ratcliffe
6385A70	American Community Survey (ACS)		
	8. ACS Applications for Time Series Methods	Tucker McElroy	Mark Asiala
	9. ACS Imputation Research and Development	Yves Thibaudeau	Beth Tyszka
	10. Data Analysis of ACS CATI-CAPI Contact History	Eric Slud	Todd Hughes
	11. Confidence Intervals for Proportions in ACS Data	Carolina Franco	Mark Asiala
	12. Mode-Based Imputation in ACS	Eric Slud	Mark Asiala
	13. Voting Rights Section 203 Model Evaluation and Enhancements Towards Mid-Decadal Determinations	Patrick Joyce	Mark Asiala
	DEMOGRAPHIC		
0906/1442	Demographic Surveys Division Special Projects		
	14. Data Integration	Ned Porter	Christopher Boniface
7523013	National Crime Victimization Survey		
	15. Contact Response Propensities and Adaptive Design Research	Isaac Dompreh	Meagan M. Meuchel
TBA	Population Division Special Projects		
	16. Introductory Sampling Workshop	Tommy Wright	Mitali Sen
7165015	Social, Economic, and Housing Statistics Division Small Area Estimation Projects		
	17. Research for Small Area Income and Poverty Estimates (SAIPE)	Jerry Maples	Wes Basel
	18. Small Area Health Insurance Estimates (SAHIE)	Ryan Janicki	Wes Basel
	ECONOMIC		
1183001	Economic Statistical Collection		
	19. Research on Imputation Methodology for the Monthly Wholesale Trade Survey	Martin Klein	Joanna Fane Lineback
	20. Use of Big Data for Retail Sales	Darcy Steeg Morris	Bill Davie
2220A10	Economic Census/Survey Engineering: Time Series Research		
	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
	Economic Census/Survey Engineering: Economic Missing Data/Product Line Data		
2220A10	Economic Census/Survey Engineering: Development/SAS		
	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	**ADMINISTRATION AND CFO**		
	27. Assessment of Finance Methodology	Tommy Wright	Robin Guinn
7236045	**CENSUS BUREAU**		
	28. National Survey of Drug Use & Health	Robert Ashmead	John Wynn

APPENDIX B

FY 2015 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 seventeenth year to provide *seven measures of performance*, this time for the fiscal year 2015. For FY 2015, the *measures of performance* for our center are:

- *Measure 1. Overall, Work Met Expectations:* Percent of FY 2015 Program Sponsored Projects/Subprojects where sponsors reported that work met their expectations.
- *Measure 2. Established Major Deadlines Met:* Percent of FY 2015 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 2015 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight.
- *Measure 3b. Plans for Implementation*: Of the FY 2015 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 2015 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 2015.
- *Measure 6. Proceedings Publications:* Number of proceedings publications documenting research that appeared in FY 2015.

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 3 thru December 11, 2015 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 2015 ***(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 2015 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 2015 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 2015 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 first two 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)
- Alisha Armas

Machine Learning & Computational Statistics Research
Bill Winkler
- Emanuel Ben-David
- Xiaoyun Lu

Statistical Computing Applications & Data Visualization Research
Martin Klein (Acting)
- Rolando Rodriguez
- Nathan Yau (FLOWING DATA.COM)

Missing Data Methods Research
Yves Thibaudeau
- Douglas Galagate (S)
- Maria Garcia
- Darcy Morris
- Jun Shao (U. of WI)

Research Computing Systems
Chad Russell
- Tom Petkunas
- Ned Porter

Simulation & Modeling Research
Martin Klein
- Isaac Dompreh
- Brett Moran

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
- David Findley
- Osbert Pang

Tucker McElroy
- James Livsey
- Aninyda Roy (UMBC)
- Thomas Trimbur

Experimentation & Modeling Research
Tommy Wright (Acting)
- Thomas Mathew (UMBC)
- Andrew Raim
- Kimberly Sellers (Georgetown U.)

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

(S) Student