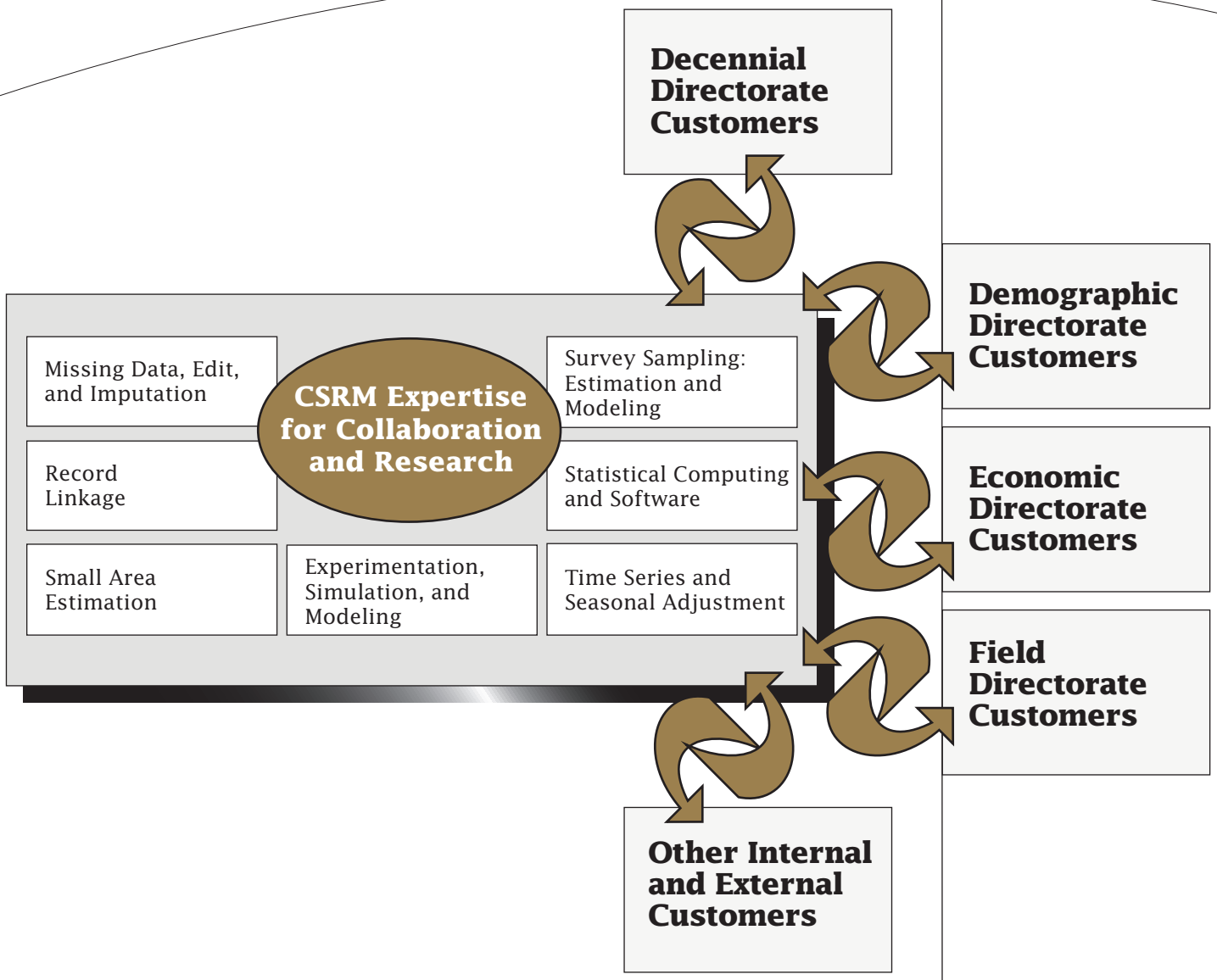
Annual Report *of the* Center for Statistical Research and Methodology

Research and Methodology Directorate

Fiscal Year 2012



United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. 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 2012, 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 2012 follow, and more details are provided within subsequent pages of this report:

- *Missing Data, Edit, and Imputation*: (1) researched and developed score functions for ranking suspicious records when using selective editing and applied to our foreign trade data; (2) developed a new application of the expectation- maximization (EM) algorithm to impute gaps in the wave of data collected by the Survey of Income and Program Participation (SIPP); and (3) made further improvements to the TEA package regarding multiple imputation methods.

- *Record Linkage*: (1) on several occasions taught the 1-day modeling/edit/imputation course; (2) developed methods for using noise multiplied data to draw inference on unknown population parameters under an assumed parametric model for the original data; (3) gave technical support to the 2020 Matching Group; and (4) updated some of the DISCRETE modeling/edit/imputation software.

- *Small Area Estimation*: (1) derived first order Taylor approximations for the replicate weight variance estimation used in the American Community Survey (ACS); (2) built a software tool for empirical evaluation and comparison of small-area models against data from an ACS-like population; and (3) initiated efforts in small area methods with misspecification and in the visualization of small area estimates.

- *Survey Sampling-Estimation and Modeling*: (1) participated in the development of numerous charts and dashboards for monitoring survey productivity and cost and in the development of discrete time hazard models for estimating response propensities in the CPS, the ACS, the SIPP, the NCVS, the NHIS, and the CES; (2) conducted research to find an automated method for detecting and treating verified influential values in the Monthly Retail Trade Survey; (3) reviewed the ranking of populations literature, conducted research, and carried out empirical studies using ACS data to compare parametric and nonparametric methods, including the bootstrap; (4) demonstrated that the method for apportioning the seats in the U.S. House of Representatives is preferred over the usual Neyman Allocation with (controlled) rounding; and (5) investigated the feasibility of social network sampling methods.

- *Statistical Computing and Software*: (1) made improvements to TEA to improve support of island area group quarters disclosure avoidance; (2) promoted activities of the R Users Group; (3) added functionality to TEA in support of research into editing and imputation for the 2020 Census; and (4) designed a Java program to extract data from .xml files as part of a web scraping feasibility investigation.

- *Time Series and Seasonal Adjustment*: (1) developed the primary alias fitter discretization method for connecting continuous time-signal extraction filters into discrete filters appropriate for stock or flow time series; (2) made minor edits and further empirical work to paper handling signal extraction for co-integrated vector time series; (3) continued empirical work on the study of HAC estimators under the context of long memory or negative memory time series; (4) continued theoretical work on the sample autocovariances and autocorrelations for linear heavy-tailed long memory time series, and generalized the consistency of the subsampling distribution estimator to the case of higher memory by imposing a smaller block size; and (5) led a spatial statistics study group.

- *Experimentation, Simulation, and Modeling:* (1) gathered data relating response propensities to demographics and geography to calibrate an existing model for allocating resources in the field; (2) tested a fiducial-based approach to handle the tolerance interval computation; and (3) identified a methodology for constructing tolerance intervals for negative binomial random variables.

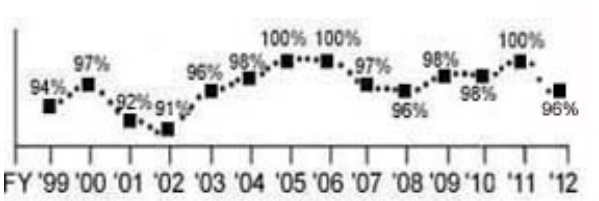
- *SUMMER AT CENSUS:* Sponsored, with divisions around the Census Bureau, scholarly, short-term visits by 26 researchers/leaders who collaborated extensively with us and presented seminars on their research. For a list of the 2012 *SUMMER AT CENSUS* scholars, see *http://www.census.gov/research/summer_at_census/ summer_2012.php*.

How Did We[1] Do...

For the 14th year, we received feedback from our sponsors. Near the end of fiscal year 2012, our efforts on 28 of our program (Decennial, Demographic, Economic, 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 14 fiscal years):

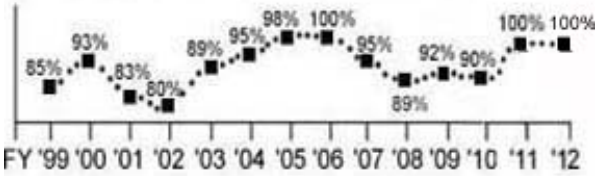
Measure 1. Overall, Work Met Expectations

Percent of FY2012 Program Sponsored Projects/Subprojects where sponsors reported that overall work met their expectations (agree or strongly agree) (27 out of 28) ……..……..... 96%



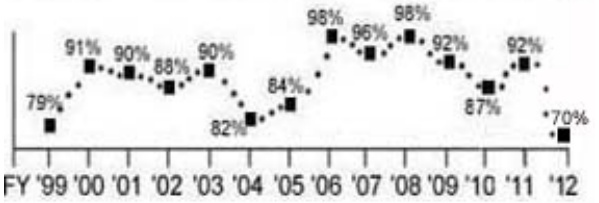
Measure 2. Established Major Deadlines Met

Percent of FY2012 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met (14 out of 14 responses) ……... 100%



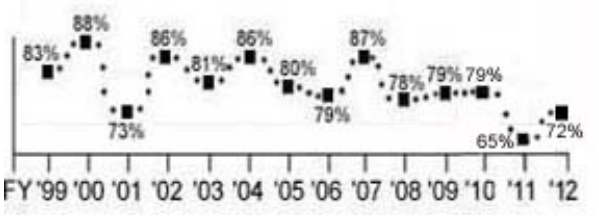
Measure 3a. At Least One Improved Method, Developed Technique , Solution, or New Insight

Percent of FY2012 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight (19 out of 27 responses) …… 70%



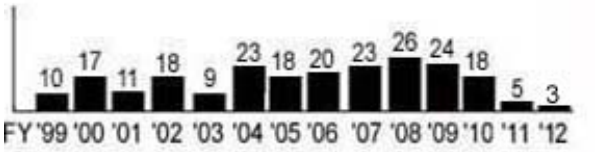
Measure 3b. Plans for Implementation

Of these FY2012 Program Sponsored Projects/Subprojects reporting at least one improved method, technique developed, solution, or new insight, the percent with plans for implementation (13 out of 18 responses) …….. 72%



Measure 4. Predict Cost Efficiencies

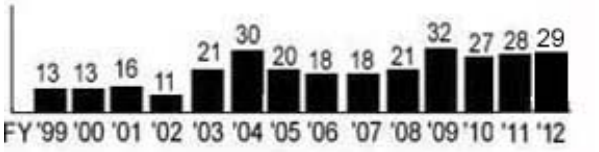
Number of FY2012 Program Sponsored Projects/Subprojects reporting at least one "predicted cost efficiency" ……..…… 3



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

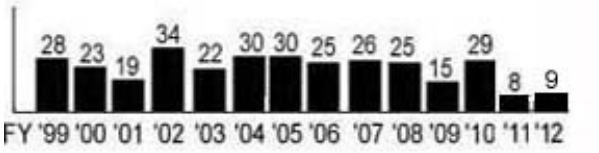
Measure 5. Journal Articles, Publications

Number of peer reviewed journal publications documenting research that appeared (15) or were accepted (14) in FY2012 ……... 29



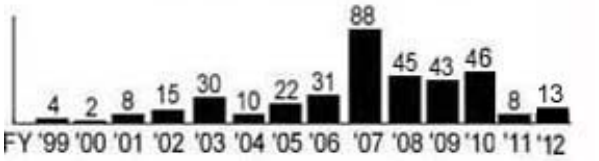
Measure 6. Proceedings, Publications

Number of proceedings publications documenting research that appeared in FY2012 ……..…. 9



Measure 7. Center Research Reports/Studies, Publications

Number of center research reports/studies publications documenting research that appeared in FY2012 ……..13



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 STATISTICAL DESIGN AND ESTIMATION (Decennial Project 5610202)

1.2 EVALUATION PLANNING COORDINATION (Decennial Project 5610206)

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: In FY 2012, staff provided background comments and papers about some of the record linkage work that the Census Bureau has done, the overlap of methods across the different areas of the Census Bureau, and how the methods can be enhanced. Staff provided both name and address standardization software and two types of record linkage software to individuals in three areas of the Economic Directorate, the Geography Division, and the Social, Economic, and Housing Statistics Division. The Director and Deputy Director suggested that all record linkage work across the Census Bureau be coordinated by the individuals in the 2020 Matching Group. Staff participated in and headed two of the three subgroups in the 2020 Matching Group that are doing research. Staff provided extensive advice (including how to construct test decks and to compare methods/software) to a subgroup that is investigating address standardization because we participated in the original research on address standardization. Staff distributed both the name and address standardizers with the Statistical Research Division (SRD) matching software. These still outperform the best commercial software according to individuals in the Center for Administrative Records Research & Applications (CARRA) and the Decennial Statistical Studies Division (DSSD).

Staff provided extensive comments to the 2020 Matching Group, particularly regarding error-rate estimation, 'real-time' matching, and determining 'true matches.' Staff provided comments on research at Carnegie-Mellon University (CMU) that is being done as part of the five-year NSF grant in support of Decennial Undercount estimation. Staff reviewed a theoretical paper on record linkage error-rate estimation by a machine learning Ph.D. student at CMU.

Staff: William Winkler (x34729), William Yancey, Ned Porter, Joshua Tokle, Michael Ikeda

B. Voting Rights Section 203 Model Based Methodology: Research, Development, and Production

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). The research undertaken sought to provide a small area model-based estimate derived from American Community Survey (ACS) 5-year data and 2010 Census data, which would provide smaller variances than ACS design-based estimates. With the production cycle having reached its completion, work on proper documentation and future research continues.

Highlights: In FY 2012, Section 203 determinations under the Voting Rights Act were released to the public. This release includes estimates and standard errors for various jurisdictions as well as various methodological and technical documentations. Staff wrote technical documentation relating the statistical methods used in the development of the estimates. Staff investigated future and alternative implementations of the work in Section 203. Staff wrote and presented a paper on the methodology at the Joint Statistical Meetings, and a journal article is in the peer review process.

Staff: Patrick Joyce (x36793), Donald Malec (NCHS), Aaron Gilary

C. Synthetic Decennial Microdata File

Description: In some cases, data users have an interest in the full microdata file whose disclosure is prohibited by law. We seek to produce synthetic individual records (microdata files) so that when we produce tables from them, the results are the same as the comparable publicly available tables from the original individual records. The synthetic individual records should be close to the underlying microdata while protecting confidentiality. The goal of this project is to produce synthetic microdata files from the decennial short form (now the American Community Survey) variables for block level geography. We are approaching the problem by using iterative proportional fitting and log linear models from fully cross-classified tables of short form variables and then creating synthetic microdata records by randomly sampling records using the estimated parameters.

Highlights: Large and sparse contingency tables naturally arise in this setting. Hence, we have considered the problem of testing homogeneity in sparse two dimensional tables. We gathered several tests that have been proposed in the literature for this problem and implemented each test in R. These tests include moment matching chi-square approximations, exact conditional tests, and tests based on an asymptotic framework

designed for the sparse data setting. We performed empirical studies to compare type I error probability and power of each test for the case of sparse binomial data. We identified specific situations in which certain otherwise adequate tests can perform poorly. We presented this work at the 2012 Joint Statistical Meetings, and a manuscript describing the work is under preparation.

Staff: Martin Klein (x37856)

D. 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:

Estimation of Synthetic Estimators in the CCM Program:

In FY 2012, staff worked with Decennial Statistical Studies Division (DSSD) staff to evaluate the uncertainty of synthetic estimators for small areas. Both frequentist and Bayesian methods were considered. Moment-based estimators had non-negligible chance of estimating negative variances, which are not valid estimates. Bayesian methods were sensitive to the number of variance groups specified. Staff is in discussions with DSSD to determine which direction to take this project.

Non-ignorable Models for Coverage Component Estimation:

Staff modified nonignorable nonresponse methods for use with survey data to perform a sensitivity analysis of the correct enumeration status for people. Staff used model diagnostics to determine appropriate link function and nonresponse model. These results are documented in the report "Effects of missing data on modeling enumeration status in the U.S. Census" by Ryan Janicki and Eric Slud (*CSRM Research Report Series*).

Coverage Measurement Planning and Development:

Staff served as consultants and reviewers for the methodology used for the coverage measurement estimates. Staff is now discussing with DSSD the research agenda for the 2020 coverage measurement program.

Uncertainty in Housing Unit Correct Enumeration Rates

Staff developed several methods to estimate the uncertainty in housing unit correct enumeration rates. Staff contributed two different methods to compute effective sample size from the coverage survey. The cell-based method, which does not borrow information from data in other areas, was picked over model-based methods for the press release of the Census Coverage Measurement results. Staff is continuing work on evaluating the model-based approaches and refining the method to estimate effective sample size (or equivalently design effects). Two conference presentations, at DC – AAPOR and Joint Statistical Meetings, were given from this work. A comprehensive report is expected in FY 2013.

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

E. Accuracy of Coverage Measurement

Description: 2010 Census Coverage Measurement (CCM) Research conducts the research necessary to develop methodology for evaluating the coverage of the 2010 Census, including new research on feasibility of triple-system estimation methods. This includes planning, designing, and conducting the research, as well as analyzing and synthesizing the results to evaluate their accuracy and quality. The focus of this research is on the design of the Census Coverage Measurement survey and estimation of components of coverage error, with a secondary emphasis on the estimation of net coverage error. Overcount and undercount estimation has not been done separately for previous censuses because of the difficulty of obtaining adequate data for unbiased estimates.

Highlights: In FY 2012, staff continued preparations to use simulation methodology to investigate the nonsampling error structure for the CCM estimates of the net census coverage error, census erroneous enumerations, and census omissions. However, this project in the 2010 Census Program for Evaluations and Experiments (CPEX) was cancelled in the second quarter of FY 2012. All materials were archived. Work did continue on the CPEX Evaluation to Assess Effect of Census Coverage Measurement (CCM) Search Area and Census Address List Formation Rules on CCM Estimates Report and staff worked closely with the author of the study report. Staff also reviewed reports for several other CPEX projects.

Staff built on knowledge of nonsampling errors in CCM to prepare a draft paper on a framework for building cost models for alternative methodologies for census taking. In particular, it provides some insight about the issues that may be encountered in the assessments of the data quality of proposed methods. Illustrative scenarios highlight how errors in the "gold standard" data set may affect the assessment of data quality (DQ) and other important data issues affecting the construction of the cost models. The concepts are demonstrated by using an enhanced CCM list as a "gold standard" with further explanation about how nonsampling errors in the CCM data may introduce errors in the enhanced CCM list.

Staff researched the sensitivity of person-level Erroneous Enumeration (EE) rate estimates in CCM to unresolved census enumerations in the post-enumeration E-sample which might be `informative' in the sense of depending

on EE status. This research examined the adequacy of the logistic regression model used in CCM and employed estimating equation methods under several alternative parametric models for the dependence between resolved-enumeration and EE status. The significant results were: (i) a complementary loglog link provided a better-fitting model than the logit link previously favored within CCM, (ii) that under the types of models considered, estimates were thoroughly insensitive within logistic regression models for resolved status allowing only a main effect for EE, and (iii) that serious and unavoidable sensitivity issues remained for EE estimates based on resolved-status models which allowed interaction terms between EE and important covariates.

Staff: Mary Mulry (x31759), Eric Slud, Ryan Janicki

F. Explaining How Census Errors Occur through Comparing Census Operations History with Census Coverage Measurement (CCM) Results
Description: The goal of this project is to understand what errors tend to be associated with the different Census operations, especially for persons and housing units removed from the census. We will compare Census files to the CCM results for a subsample of CCM areas. This comparison is intended to help find patterns of errors in Census operations and provide insights into ways to avoid these errors.

Highlights: In FY 2012, staff produced a working draft of the software specification for the person computer matching and processing. Staff also provided comments on a draft clerical matching specification produced by the procedures and training branch of the Decennial Statistical Studies Division. Following the cancellation of the evaluation, staff completed closeout activities for the evaluation. As part of these activities, staff created a draft memorandum documenting the universe counts for the invalid/valid census matching universe and a second draft memorandum documenting the variables on the universe file for the invalid/valid census matching universe. Draft specifications and other documents related to the evaluation have been given to the Decennial area. Work is considered complete for this project.

Staff: Michael Ikeda (x31756), Mary Mulry

1.3 CODING, EDITING, AND IMPUTATION STUDY
(Decennial Project 6510201)

A. Software Development (TEA)
Description: Here we report applications of TEA software to the coming 2020 Census. For the broader project, see General Research 0351000 and 1871000, *Statistical and Computing Software* (A) TEA Software Development.

Highlights: In FY 2012, staff added functionality to their R package, TEA, in support of research into editing and imputation for the 2020 Decennial Census. New functionality includes group-level and nested recoding, classification and regression tree modeling, and sequential regression modeling. Staff has added documentation and has incorporated testing procedures to improve usability and dependability of TEA.

Staff: Ben Klemens (x36864), Rolando Rodriguez, Yves Thibaudeau

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: In FY 2012, staff wrote code which implements imputation methods (specifically, sequential regression and regression trees) to be used as contenders against the current hot deck. Staff added additional code to increase production infrastructure, and this code improves support for group-level recodes and nested recodes. Staff developed demonstration code for the R-language library, TEA, that holds the code and has written documentation in the form of R vignettes. Staff completed preparation to begin testing imputations methods. Testing of these methods will commence upon receipt of test data.

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

1.4 ENHANCING DEMOGRAPHIC ANALYSIS
(Decennial Project 6710201)

A. Enhancing Demographic Analysis for the 2020 Census
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: In FY 2012, staff began the work of planning the methodology for the development of an agenda for future utilization. *S*taff researched methods of variance estimation for population projections and limitations on existing scenario-based methods.

Staff: Ben Klemens (x36864)

1.5 PRIVACY AND CONFIDENTIALITY STUDY (Decennial Project 6810204)

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: In FY 2012, staff participated in planning of the design and scope of this project, and in the development of the Privacy and Confidentiality Project Plan. Staff participated in planning for an attitudinal CATI follow-up survey of responders and nonresponders aimed at exploring attitudes regarding contact modes and administrative records. This project is coordinated by the 2020 Census Privacy and Confidentiality Team. Staff participated in team meetings with the scientific panel; as a result of these meetings, the project plan was revised and the team was restructured.

Staff: Martin Klein (x37856)

1.6 MATCHING PROCESS IMPROVEMENT (Decennial Project 6810205)

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: In FY 2012, staff participated in the planning and preparation process for the research, particularly in the literature review, data needs, the risk register, and identification of medium-term research projects. Team documents have been updated for the 30-day, 60-day, 90-day, and 120-day reviews. Staff has been assigned to the error-rate estimation and real-time matching sub-teams and to medium-term research projects on investigating duplication within the 2010 Census (using CCM clerical as "truth") and on defining geographic distance in matching. Staff modified the matching system used in previous research for use on the 2010 Census Unedited File (CUF) and performed a matching of the 2010 CUF against itself and began analyzing the results. The results so far appear generally consistent with the 2010 Duplicate Person Identification results and the basic patterns appear fairly similar to the corresponding research results based on the 2000 Census, although there are fewer links with matching phone number than in 2000 and the proportion of within-block links is somewhat lower than in 2000. One observation on patterns of geographic distance is that the distances do separate out reasonably well by the current geographic categories (block, tract, county, state, nation) although the tails of the distance distribution for each category are worth further examination. Also, about one-sixth of the links have missing distance due to missing latitude or longitude for at least one of the units. Links with missing distance are more likely than links with non-missing distance to be within-block or within-tract links. Staff provided advice on how to perform 'real-time' matching which is needed for some of the 2020 projects. In particular, our *BigMatch* methods/software which were used for 2010 Decennial Census production matching were 'real-time'. The type of hardware that is currently used is suitable for 'real-time' matching in a number of large situations. In other smaller situations modifications of some of the record linkage methods used for the 1990 Housing Unit Coverage Study or the 1992 Census of Agriculture list development (12 lists with 16 million records) would likewise be suitable. Staff provided extensive advice on how to construct test decks and how to compare methods/systems for standardization or record linkage when there are no suitable test decks during initial tests. Staff also provided extensive advice about record linkage error-rate estimation (including how it was performed in the past for decennial applications, certain smaller administrative-records tests by DSSD, and for matching applications in the Economic Directorate) and provided pointers to the literature on how statistical analyses are adjusted for matching error. Staff twice taught a 2-day record linkage course and also taught a 1-day record linkage error-rate course, all to Census Bureau personnel.

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

1.7 STATISTICAL DESIGN FOR 2020 PLANNING, EXPERIMENTATION, AND EVALUATIONS (Decennial Project TBA)

A. Modeling Successive Enumerator Contacts for Nonresponse Followup (NRFU)

Description: One facet of the NRFU operations analysis includes the number of contact attempts made and the impact on data quality. Current summaries of these data include the distribution of the number of contacts and the

distribution of the final mode of contact. This project aims to provide additional assessments of this data, including the distribution of waiting times between successive contacts as well as the effect of these times on the eventual outcome. The intent of any additional analysis is to provide guidance regarding possible designed experiments in anticipation of the 2020 Decennial Census.

Highlights: In FY 2012, staff obtained relevant NRFU data and has started understanding the data structure. The 2020 Planning Team expressed interest in this effort to incorporate with their research about reducing the number of contacts to reduce costs. Discussions with staff in DSSD have involved preliminary contact research.

Staff: Derek Young (x36347)

B. 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 2012, staff used logistic and linear regression models from a previous TAC assessment as a starting point for modeling efforts with the MAF. We then identified a stepwise framework to characterize sources of coverage errors in the MAF. We also identified potential independent variables for our eventual error model, including the complexity of the block, an indicator for small multi-units, and the incorporation of the American Community Survey (ACS) HU and POP count results (before weighting). Efforts have been made with the Geography Division (GEO) regarding validation of the MAF error model and potential integration with the Geographic Support System Initiative (GSSI). We also conducted a literature review that highlights previous methodology regarding undercoverage and overcoverage calculations. Data access requests for staff are currently being processed.

Staff: Derek Young (x36347)

C. 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 followup to significantly reduce decennial census cost while maintaining quality. The project is coordinated by one of the 2020 Census Integrated Project Teams.

Highlights: In FY 2012, staff was assigned to the Cost Modeling and Simulation sub-teams and participated in the planning and preparation process for the research, particularly in the Cost Modeling sub-team's data needs, literature review, Scientific Review Information Sheet, workload management requirements, goals/objectives for upcoming field tests, and Study Plan. Team and sub-team documents have been updated for the 30-day, 60-day, 90-day, and 120-day reviews. The Cost Modeling sub-team's Scientific Review Information Sheet and Study Plan are both in near final form. The study plan has been sent to outside reviewers for comments. The Simulation sub-team has released two memoranda for the record: "2020 Census 8.107 Supplementing and Supporting Nonresponse with Administrative Records (SSNAR) Simulation Subteam 2010 Census NRFU Model Evaluation Template--Version #1" (presenting a set of evaluation criteria for measuring accuracy of 2010 Census NRFU simulation models), and "2020 Census 8.107 Supplementing and Supporting Nonresponse with Administrative Records (SSNAR) Simulation Subteam 2010 Census NRFU Unit-Level Cost Methodology & Documentation--Version #1" (documents methodology for producing six housing unit level cost measures for the 2010 Census NRFU operation). Sub-team members reviewed both memoranda before release.

Staff: Michael Ikeda (x31756), Ned Porter

D. Local Update of Census Addresses (LUCA) Program Improvement

Description: The purpose of this project is to assess all facets of the LUCA program to identify cost effective changes, improve the quality of the Master Address File, and optimize the benefits derived by the Census Bureau and the participants. The project is coordinated by one of the 2020 Census Integrated Project Teams.

Highlights: In FY 2012, staff participated in the planning and participation process for the research. Team documents were updated for the 30-day, 60-day, 90-day, and 120-day reviews. The record linkage error-estimation course mentioned in the 2020 unduplication research section (1.6A) also includes concepts applicable to this research.

Staff: Michael Ikeda (x31756), Ned Porter

1.8 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 5385260)

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: In FY 2012, staff continued work on re-centering ACS multi-year estimates. The method for re-weighting based on a sampled Brownian Motion plus noise model was implemented and tested, and an estimate of the signal-to-noise ratio was constructed. The method's properties were explained and demonstrated to the sponsors. Staff also continued making revisions to a paper that explores the comparability of multi-year estimates with linear estimates.

Staff: Tucker McElroy (x33227)

B. ACS Variances

Description: Work under this heading this year concerned four research projects: (i) simultaneous nonresponse adjustment, calibration, and weight trimming with new examples on Survey of Income and Program Participation (SIPP) data; (ii) comparison of alternative methods of estimating variances for complex survey estimates with a new method, 'hybrid'+ estimates of cross-classified population totals containing design-based cell size estimates for some cross-classified categories, and model-based conditional-probability estimates within other categories; (iii) performance of (Fay-method) BRR variance estimation under misspecified calibration or raking-ratio adjustment for nonresponse; and (iv) (Slud only) development and comparison of methods to assess the validity of 0 estimates in ACS tables. [See Decennial Project 5610202 (E).]

Highlights: In FY 2012, staff revised a journal paper on topic (i). Staff gave a talk as Discussant in a Demographic Statistical Methods Division (DSMD) seminar on topic (ii) and worked on revising an earlier report on this research into a *CSRM Research Report* and journal submission. Staff prepared an ASA abstract on topic (iii) in these quarters and proposed using analogous methods in the Coverage Measurement estimates of Erroneous Enumeration rates for Housing Units coverage work. The methodology of topic (iv) initiated in previous quarters on variance-calculations and prediction intervals to assess the validity of 0 estimates in ACS tables was applied during this period to the Small-Area modeling of Census Coverage Measurement estimates of Erroneous Enumeration rates for Housing Units under section 1.1.E. Work is complete on this project.

Staff: Eric Slud (x34991), Yves Thibaudeau

C. ACS Data Issues

Description: Various issues related to the quality and presentation of ACS estimates were discussed and investigated by small interdivisional teams or division staff. The goal of these investigations was to make recommendations to aid in resolving the issues.

Highlights: In FY 2012, the project report was reviewed and finalized. It is available as *CSRM* and *ACS Research Reports*. A description of the project was given to the ACS Research Group. Work is complete on this project.

Staff: Lynn Weidman (x34902), Julie Tsay, Michael Ikeda

D. ACS Exploratory Analysis of the Differences in ACS Respondent Characteristics between the Mandatory and Voluntary Response Methods

Description: In 2002 and 2003, the Census Bureau conducted research to determine whether the ACS could be implemented as a voluntary, rather than a mandatory, sample survey. It showed a decrease of over twenty percent in mail response when the survey was voluntary, and the reliability of estimates was adversely impacted by the reduction in the total number of completed interviews. This project revisits the test data and uses logit models to determine person, housing unit, and household characteristics most closely related to lower response under voluntary treatment. Also studied is whether their corresponding estimates at the national level are significantly different. This project revisits the test data and uses logit models to determine person and housing unit/household characteristics most closely related to differential response between the voluntary and mandatory methods. Also studied is whether related profile estimates at the national level for the two methods are significantly different.

Highlights: The project report was reviewed and finalized. It is available as *CSRM* and *ACS Research Reports*. A description of the project was given to the ACS Research Group. Respondent characteristics that differed nationally between the mandatory and voluntary methods and also showed differences between estimates from the methods included household type, educational attainment, moved or not in last year, and language spoken at home. Work is complete on this project.

Staff: Lynn Weidman (x34902), Michael Ikeda, Julie Tsay

E. ACS Imputation Research and Development

Description: The American Community Survey 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 to quality problems remaining in the micro-data. Before the data are ready for public release, they have 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: In FY 2012, staff presented a plan to upgrade the editing and imputation software functions as part of the processing of the American Community Survey data. Staff proposed adapting the software TEA, developed by center staff. Staff is preparing a comprehensive strategy, consulting with the American Community Service Office (ACSO) and the Social, Economic, and Housing Statistics Division (SEHSD) to coordinate the development of the TEA prototype with support from data reviewers in the Demographic Directorate. Staff presented a plan for "Simultaneous Editing in Near-Real Time for ACS Data" to the ACS Research & Evaluation Team, and to the ACS Research and Evaluation Steering Committee.

The simultaneous-edit model, when successfully implemented, eliminates the need for lengthy and costly manual reviews of the data. In addition, in the context of a modern computing environment, a simultaneous-editing process can be implemented in "real-time". There are numerous potential payoffs from real time processing. They include better quality control and better survey cost management. In that context we developed imputation methods that run in real time and lead to estimators of survey indicators that integrate imputed responses and of their variance. One of the most promising methods we have developed is the random multiple-imputation hot-deck. It is now available in "TEA," a general software for processing surveys developed by CSRM. The random hot-deck runs at least one order of magnitude faster than traditional model-based multiple imputation (Garcia, Erdman, Klemens 2012). It can be implemented in real-time (24 hours). The multiple imputation features of the random hot-deck makes possible the estimation of the variance of survey indicators that integrate imputed responses. These indicators and their variances enable real time decisions for optimizing field operations. The random multiple-imputation hot-deck will be tested as part of the adaptive design approach to ACS data collection. We plan on using the Group Quarters data for an operational test.

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

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

A. Tobacco Use Supplement (NCI) Small Domain Models

Description: Staff is working with Demographic Statistical Methods Division (DSMD) on a project for the National Cancer Institute (NCI), studying the relationship between smoking status and a range of geographic/ demographic covariates. Using the Tobacco Use Supplement to the Current Population Survey (TUS-CPS), staff is assisting NCI toward making estimates of smoking related behavior using county-level or state-level dependent variables (e.g., percent of males, percent Hispanic, percent below poverty level). The goal is to identify where anti-smoking funds could best be directed.

Highlights: In FY 2012, staff worked with Benmei Liu (National Institutes of Health) to develop models for smoking-related rate variables (e.g., former smoking, people banned on smoking at work, etc.) for U.S. counties. Staff developed the plan of modeling each of these rates using stepwise regression modeling, and using these model-based estimates as input to a Fay-Herriot small-area modeling for estimating those variables, to be implemented using Gibbs Sampling software. Together with Dr. Liu, staff wrote the programming for each of these models to estimate the rate of people currently smoking, and performed the modeling on weight-adjusted CPS data to generate the small area estimates, incorporating features of the sampling scheme into the model. Staff honed the modeling with potential covariates from Census 2000 and subsequently adjusted the model for potential covariates taken from the 2006-2007 estimates of various surveys, which are more contemporaneous to the CPS data. Currently, staff is producing diagnostics to evaluate and explore modifications to the modeling, and the plan is eventually to produce and release estimates for each variable for all U.S. counties.

Staff: Aaron Gilary (x39660), Partha Lahiri

B. 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 also the possibility of altering CPS weighting procedures to allow for a single simultaneous stage of weight-adjustment for nonresponse and population controls.

Highlights: In FY 2012, staff reviewed CPS documentation and historical research papers, related them to sample survey theory, and explored possible alterations in design parameters and composite estimators. Then staff initiated discussion of the feasibility of applying a previously developed method of simultaneous regularity with CPS Branch nonresponse adjustment, calibration, and weight compression, to CPS.

Staff: Eric Slud (x34991), Yang Cheng (DSMD), Reid Rottach (DSMD).

1.10 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: In FY 2012, staff submitted the paper *Exploring re-identification risks in public domains* by Aditi Ramachandran, Lisa Singh, Ned Porter and Frank Nagle to the conference Privacy, Security, and Trust 2012. The paper was accepted and presented by Aditi Ramachandran at the conference in Paris. The paper was published in the *CSRM Research Report Series* as well. This phase of the project is now complete. A new approach enhancing Public Use Microdata Files through published tabular data has been proposed by the Center for Disclosure Avoidance Research (CDAR), and staff will experiment with this idea.

Staff: Ned Porter (x31798), Lisa Singh (CDAR), Rolando Rodríguez

1.11 NATIONAL CRIME VICTIMIZATION SURVEY (Demographic Project 7323008/7523012)

A. Analyzing the Effects of Sample Reinstatement, Refresher Training Experiment, and Process Monitoring and Fitness for Use

Analyzing the Effects of Sample Reinstatement

Description: During 2010 and 2011, the National Crime Victimization Survey sample size was restored (increased) to previous levels. This, in conjunction with the realignment imposed by the closing of six Regional Offices, brought changes to interviewer workloads, with possible impact on victimization measures for households and persons. Through analysis of survey outcomes and paradata, we seek to quantify the effects of reinstatement and realignment on victimization rates.

Highlights: In collaboration with staff from the Victimization and Expenditure Branch of the Demographic Statistical Methods Division (DSMD), in FY 2012, we fit a variety of marginal regression models using generalized estimating equations (GEE) to longitudinal data on interviewer productivity rates, where a productivity rate is defined as the expected number of victimizations among households or persons scaled by the interviewer workload.

We discovered that productivity decreased during and after the period of reinstatement. In a second round of modeling, we attempted to separate changes in productivity into components due to change in the rates of successful interview and rates of reported victimization among the successful interviews. We hypothesized that much of the change would be confined to the interview rates. That hypothesis proved to be false; the changes in productivity could not be explained by changes in interview rates.

In response to input from the Bureau of Justice Statistics, in the third quarter of FY 2012, we formulated a flexible class of Bayesian models. The outcome, a measure of workload per interviewer, is described by a generalized linear regression model with a smoothly varying time trend captured by a penalized natural cubic spline. The splines are allowed to vary with nested random effects due to interviewers and regional offices. We are developing efficient, custom software to fit these models via Markov Chain Monte Carlo.

Refresher Training Experiment

Description: In 2011, an experiment was embedded within the NCVS. Teams of interviewers were randomly assigned to two cohorts. The first cohort received specialized training that was designed to improve the quality of the interview process, and the second cohort received the same training six months later. Through modeling of survey outcomes and paradata, we seek to quantify the effects of the so-called Refresher Training program on victimization rates.

Highlights: In FY 2012, the experiment was conducted as a matched-pair cluster randomized trial, with teams of interviewers forming the clusters. To account for this experimental design and for correlations among repeated measurements from the same interviewer, we fit a sequence of multilevel Poisson regression models with random effects for interviewers and team pairs. Parameters were estimated by Bayesian Markov Chain Monte Carlo. We estimated that Refresher Training increased the rates of household victimization by about 30%, and this increase is statistically significant. Refresher Training had no appreciable effects on person victimizations.

The models for Refresher Training were refit using data from the entire 2011 calendar year. Training appears to have increased household victimizations by about 37 percent on a multiplicative scale. No significant effect was seen on person victimizations.

Process Monitoring and Fitness for Use

Description: Information gathered from NCVS field operations is synthesized into variables that serve as indicators of data quality. In this project, we are developing classes of flexible models and graphical tools

for describing how these variables evolve over time. These techniques are intended to help the Census Bureau and the Bureau of Justice Statistics staff to monitor the performance of field staff, to describe the effects of interventions on the data collection process, to quickly alert survey management to unexpected developments that that may require remedial action, and to assess the overall quality of NCVS data and their fitness for use.

Highlights: In FY 2012, drawing upon the vast literature on semiparametric regression, we developed techniques for Bayesian estimation and prediction based on penalized splines cast as generalized linear mixed models. We are now extending these models to describe paradata series for individual interviewers. Models for paradata series from individual interviewers were formulated under the flexible class of Bayesian models created in response to input from the Bureau of Justice Statistics. Software for fitting these models is being developed in R with native routines written in Fortran. The results are being summarized for presentation at a workshop in September.

Staff: Joe Schafer (x31823)

1.12 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: In FY 21012, staff developed and encoded a method to make forecasts of age-specific fertility and mortality rates, stratified by age and racial groups. The method involved dimension reduction and time series modeling, based on earlier work by William Bell. Final results were exhibited to the clients and the R code was demonstrated to them. The clients indicated they would use the program to generate their final projections.

Staff: Tucker McElroy (x33227), Osbert Pang, William Bell (R&M).

1.13 SURVEY OF INCOME AND PROGRAM PARTICIPATION IMPROVEMENTS RESEARCH (Demographic Project 1465444)

A. Model-Based Imputation for the Demographic Directorate
Description: Staff has been asked to review and improve ultimately all of the imputation methodology in demographic surveys, beginning with the Survey of Income and Program Participation and the Current Population Survey.

Highlights: In FY 2012, staff worked on two separate model-based imputation methods and associated computer programs (implemented SAS and R respectively) for imputing monthly earning at the job level in the Survey of Income and Program Participation (SIPP). We designed a simulation study to compare these two models. Starting with a set of completed SIPP data, we created 25 SIPP data files with missing at random monthly earnings. We ran the two computer programs to produce four multiply-imputed data sets for each of the 25 simulated data files. Results of the evaluation study showed that for each month, the R program performs significantly better than the SAS program provided by the Social, Economic, and Housing Statistics Division (SEHSD) in terms of Root Mean Squared Error (RMSE.)

We also completed an additional evaluation of the simulation study in which we compared the variances in the estimate of the means of monthly SIPP earnings. We used Rubin's formulae (1987) to compute within-imputation, between-imputation and the total estimated variances due to imputation. Results of this evaluation study showed that for each variable and each type of variance, the CSRM model performs better or there is no significant difference than the SEHSD model.

We designed an additional simulation study to compare the performance of model-based imputation methods with the hot-deck method which is the methodology currently used for SIPP imputation. For this simulation study, we begin with the same completed SIPP file, and randomly select 10 percent of observations to be set to missing to create 100 replicates with missing at random monthly earnings and earnings indicators. We multiply-impute using the R computer program and the hot-deck model as implemented in CSRM generalized edit and imputation system TEA. We found that, in each month, the model-based imputation method significantly outperforms the hot-deck procedure in terms of RMSE.

Staff wrote a draft for a *CSRM Research Report* describing the results of this study "A Comparison of Multiple Imputation Methods for Imputing Earnings in the Survey of Income and Program Participation."

Staff: Maria Garcia (x31703), Chandra Erdman, Ben Klemens, Yves Thibaudeau

1.14 SOCIAL, ECONOMIC, AND HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165000)

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: In FY 2012, staff compared several bivariate models using the Census 2000 long-form, previous year American Community Survey (ACS) and the previous 5-year ACS data as the second equation. The evaluation showed that using multi-year ACS 2005-2009 estimates was the preferred model than using the outdated Census 2000 or the previous ACS 2009 estimates in the bivariate model. These results are documented in "An Empirical Study on Using Previous American Community Survey Data Versus Census 2000 Data in Models for SAIPE Poverty Estimates" by Elizabeth T. Huang and William R. Bell in the *CSRM Research Report Series*. Staff implemented a new small area estimation model, using the zero-inflated beta distribution as the sampling model and a random effects logistic regression as the linking model. The model is intended to improve on the current county-level SAIPE production model, which cannot currently use counties with ACS direct estimates of 0 when fitting the model. Simulations using the ACS simulation study software suggest the zero-inflated beta model improves on the production model in several regards for small counties. Staff also developed a small area model for the design-based variance of the poverty rate estimates from the ACS. To specific the uncertainty of the design-based variance, a first-order Taylor approximation was used. Staff gave several presentations of these model frameworks at the Joint Statistical Meetings and seminars within the Census Bureau.

Staff: Jerry Maples (x32873), Jerzy Wieczorek, William Bell (R&M)

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.

Highlights: In FY 2012, staff extended previous work on estimating parameters at two different geographic levels when there are constraints on the parameters. The methods developed differ from previous approaches in that benchmarking constraints are incorporated into the distinct small area models rather than point estimates. Two new approaches, one based on minimum discrimination information, in which the first two moments are constrained, and one based on the full conditional distributions, were introduced. This work was extended to the case of parameter estimation at three different geographic levels for the special case where the constraints on the parameters are linear and the loss function is quadratic. A draft of a paper documenting this work has been written.

Staff: Ryan Janicki (x35725)

1.15 IMPROVING POVERTY MEASURES/IOE (Demographic Project 1442555)

A. Tract Level 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: In FY 2012, staff attended planning meetings and is currently assembling a team while our partners in the SEHSD are preparing the datasets to be used for the research project. Staff have been reviewing proposed county level models to determine the suitability for tract level estimates.

Staff: Jerry Maples (x32873), Elizabeth Huang, Ryan Janicki, Aaron Gilary, William Bell (R&M)

B. 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: In FY 2012, staff attended planning meetings and currently working to identify additional personnel for this project.

Staff: Jerry Maples (x32873), Ryan Janicki, Aaron Gilary, Jerzy Wieczorek, William Bell (R&M)

1.16 EDITING METHODS DEVELOPMENT (Economic Project 2320254)

A. Investigation of Selective Editing Procedures for Foreign Trade Programs
Description: The purpose of this project is to develop selective editing strategies for the U.S. Census Bureau foreign trade statistics program. The Foreign Trade Division (FTD) processes more than five million transaction records every month using a parameter file called the Edit Master. In this project, we investigate the feasibility of using selective editing for identifying the most erroneous records without the use of parameters.

Highlights: In FY 2012, staff researched and developed score functions for selective editing of our foreign trade data. Selective editing requires a score function to assign a priority ranking to erroneous records. We proposed a score function that includes a measure of the probability that the record is suspicious and a measure of the relative effect errors in the suspicious record have on the macro estimates (totals).

We completed a feasibility study and used the pseudo-bias as defined by Latouche and Berthelot (1992) to assess the impact of a selective editing application to these data. Results showed the slope of the pseudo-bias rapidly decreasing as the highly suspicious records are reviewed and the selective editing total approaches the final publication total. We described the score functions and presented results of the feasibility study in a *CSRM Research Report* ("Score Functions for Selective Editing of the US Census Bureau Foreign Trade Data.")

We designed an evaluation study following on the recommendations of the feasibility study. We updated the selective editing program taking into account the availability of more data to calculate the parameters needed to assign measures of suspicion and potential impact. We calculated hit rates to compare the final edited data with the selective editing final data. For most commodity groupings, the proportion of records flagged for follow-up that are true rejects was at most 65%. Upon further analysis, we concluded hit rates are not a good evaluation measure of the effectiveness of selective editing; we are not attempting to re-engineer the current editing procedures so it's possible a record ranked as highly suspicious by the selective editing program was not "hit" by the Edit Master and vice versa. We concluded that the pseudo-bias is still the best way to measure a selective editing application to these data.

Staff: Maria Garcia (x31703), Yves Thibaudeau, Andreana Able (FTD)

1.17 TIME SERIES RESEARCH (Economic Project 2320252)

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 and benchmarking software used by the Economic Directorate.

Highlights; In FY 2012, staff provided seasonal adjustment and X-12-ARIMA support to the following: J. P. Morgan, FP & L Utilities, Bank of England, SAS Corporation, World Business Chicago, Sanford C. Bernstein & Co., Anheuser-Busch, Capgemini, RBC Global Asset Management (US), HRL Morrison, Ford Motor Company, Toyota (Brazil), Itau-BBA (Brazil), True North Management Group, Rosen Consulting Group, Archstone Consulting, RBC, Trulia, Office fédéral de la statistique OFS (Switzerland), Instituto Brasileiro de Geografia e Estatística (Brazil), International Monetary Fund, Oesterreichische Nationalbank (Austria), CitiBank, Bank of India, Bank of England, Bank of Japan, National Bank of Belgium, Bundesbank, Banco Central do Brasil, Colorado Department of Labor, Bureau of Economic Analysis, Bureau of Labor Statistics, Department of Transportation, National Bureau of Statistics of China, Office of National Statistics (UK), Instituto Nacional de Estadística (Spain), Statistics Australia, Australian Bureau of Statistics, Statistics New Zealand, Statistics Sweden, National Institute of Banking and Finance (Pakistan), Statistics Netherlands, Statistics South Africa, University of Warwick, Drexel University, University of Manitoba, University of Mannheim, China Academy of Transportation Sciences.

Staff met several times with analysts from the Economic Directorate to discuss how the recent recession has affected seasonal adjustments at the Census Bureau and prepare for meetings with the Chief Economist of the Commerce Department and the Chief Economist of the Labor Department. Our staff also worked with staff from the Economic Directorate to develop a response to an email from a senior statistician at the Bundesbank relating to the future of seasonal adjustment software at the Census Bureau, which was shared with other statistical agencies in Europe.

Staff spoke with analysts from the Bureau of Labor Statistics concerning the usefulness of spectral plots for quarterly series, and organized meetings with analysts

from the Economic and Social Research Institute (Japan), the Australian Bureau of Statistics, and Statistics Sweden to discuss issues related to time series analysis and seasonal adjustment.

Staff: Brian Monsell (x31721), Tucker McElroy, Christopher Blakely, Osbert Pang, David Findley (Consultant)

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. This fiscal year's goals include: (1) finishing a version of the X-13ARIMA-SEATS program with accessible output and improved performance so that, when appropriate, SEATS adjustments can be produced by the Economic Directorate; (2) developing software system that provides a simulation environment for X-13 seasonal adjustments called USim-X13; and (3) incorporating further improvements to the X-12-ARIMA/X-13A-S user interface, output and documentation. In coordination and collaboration with the Time Series Methods Staff of the Office of Statistical Methods and Research for Economic Programs (OSMREP), the staff will provide internal and/or external training in the use of X-12-ARIMA and the associated programs, such as X-12-Graph, when appropriate.

Highlights: In FY 2012, staff released an updated version of X-13ARIMA-SEATS (Build 148), to the Economic Directorate for their testing and then released a revision (Build 149) to the general public in July of 2012, which included incorporating an index of tables directly into the output file of the accessible version of the software. In addition, comments from the Time Series Methods Staff were incorporated into the software. Staff compared adjustments from this version of the software to the last released version of X-12-ARIMA (Build 192) and found in most cases no differences in the adjustments, and small differences between compilers on Linux-based machines. Staff also provided support for analysts and programmers in the Economic Directorate in their testing of the new release. Staff repaired a defect in X-13ARIMA-SEATS to ensure diagnostics produced by X-13ARIMA-SEATS in the automatic model identification output is generated using the proper residuals.

Staff developed versions of the X-13ARIMA-SEATS prototype to incorporate outlier effects used in research on modeling recession effects as well as updated versions of the SEATS model-based seasonal adjustment software. Staff also updated the latest version of the X-12-ARIMA source code with links to the FAME database for further development by Statistics Australia.

Staff worked to incorporate the most recent version of X-13ARIMA-SEATS into iMetrica (formerly known as uSim-X-13) and implemented a mouse-on-canvas approach to computing seasonal adjustment in blocks. New modules were developed that perform Bayesian computation of general ARIMA (with fractional differencing) models as well as a few nonGaussian models, modules which estimate two different versions of dynamic factor models. In addition, an interface to utilize signal extraction filters computed in SEATS in the direct filter approach module was developed, which allows for real-time customized estimates of seasonal adjustment, trend adjustment, and trend-irregular adjustment. Furthermore, a new adaptive approach for handling direct filtering of univariate time series has been developed in which enables an even faster allocation real-time detection of turning points in economic data. An interface in the direct filter MDFA module has been developed to engage users in the control of this new adaptive approach. Finally, a sliding spans analysis tool has been added to the uSimX13 module for performing SARIMA modeling and signal extraction on different date spans.

Staff: Brian Monsell (x31721), Christopher Blakely, 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 is 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; (9) studying nonparametric methods for estimating regression effects, and their behavior under long range dependence and/or extreme values.

Highlights: In FY 2012, staff worked on a number of projects which included: (a) completed empirical work and documentation for a project allowing seasonal adjustment of mixed frequency time series data observed as a stock or a flow; (b) continued implementation and simulation studies for a multiple-testing framework for Ljung-Box diagnostic statistics; (c) derived the class of semi-group operators that correspond to data transformations such that the transforms are homomorphisms, with the application of understanding seasonal adjustment decompositions appropriately; (d) investigated a multivariate seasonal adjustment procedure based on using a dynamic factor model to first reduce dimension. Code was written and tested on several dozen monthly time series. (e) Staff continued research in modeling moving holidays in stock series, by comparing the smoothness of adjustments from flow and stock holiday adjustments. (f) Staff investigated non-orthogonal component decompositions and their ramifications on seasonal adjustment methodologies

Staff: Tucker McElroy (x33227), Christopher Blakely, Brian Monsell, Osbert Pang, William Bell (Research and Methodology Directorate), David Findley (Consultant)

D. Supporting Documentation and Software for X-12-ARIMA and X-13A-S

Description: The purpose of this project is to develop supplementary documentation and utilities for X-12-ARIMA and X-13A-S 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, rendering the output from X-13A-S accessible, further developing the Usim-X13 software and documentation, and exploring the use of component and Java software developed at the National Bank of Belgium.

Highlights: In FY 2012, staff released an updated version of the *Genhol* utility to the public. Staff updated documentation for the X-13-ARIMA-SEATS program to prepare the software for public release, including developing documentation for the version of the software that generates accessible HTML output. Staff began developing documentation for the different modules of the *iMetrica* (formerly uSim-X-13) software. Staff made a version of the RegCMPNT program available to interested researchers in the US and the Netherlands.

Maintenance of the X-12-ARIMA and X-13ARIMA-SEATS websites continued to ensure that they follow standards established by the Census Bureau and to improve search engine performance. Staff also added several papers to the Seasonal Adjustment Papers website, and updated the utility used to generate the code for that website.

Staff developed training materials and examples for a seasonal adjustment class taught in Beijing, China using Genhol, X-13ARIMA-SEATS, and other software.

Staff: Brian Monsell (x31721), Tucker McElroy, Christopher Blakely, David Findley (Consultant)

1.18 GOVERNMENTS DIVISION PROJECT ON DECISION-BASED ESTIMATION (Economic Project TBA)

Description: This project involves providing consultative work for Governments Division on point and variance estimation for total government employment and payrolls in the Survey of Public Employment and Payroll, within a 'decision-based' method stratumwise GREG estimation after collapsing substrata of small versus large units according to the results of hypothesis tests on equality of regression slopes.

Highlights: In FY 2012, staff completed and submitted a journal paper (joint with J. Shao, Y. Cheng, S. Wang, and C. Hogue) on variance estimation for decision-based survey estimators, comparing substitution (inclusion-probability formula) estimators against bootstrap method, and are now revising the paper. Additionally, staff consulted with Governments Division on methodology related to the use of these methods in Governments surveys, and proposed, based on preliminary data analysis, to improve the state-by-government type estimates of total payroll and numbers of employees by using additional regressors in GREG estimation. Work is complete on this project.

Staff: Eric Slud (x34991), Bac Tran (GOVS), Gauri Datta

1.19 PROGRAM DIVISION OVERHEAD (Census Bureau Project 0381000)

A. Center Leadership and Support

This staff provides ongoing leadership and support for the overall collaborative consulting, research, and operation of the center.

Staff: Tommy Wright (x31702), Alisha Armas, Michael Hawkins, Michael Leibert, Erica Magruder, Gloria Prout, Joe Schafer, Eric Slud, Esan Sumner, Kelly Taylor, Sarah Wilson

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: In FY 2012, staff from the Center for Statistical Research and Methodology (CSRM) and the Center for Economic Studies (CES), along with external researchers Lars Vilhuber and Kevin McKinney, began working on a project to design a high performance research computing environment that would better serve the Census Bureau research community. The project is being sponsored and managed by the IT Directorate, with the goal of addressing problems with the current environment. A two-day workshop to identify the functional requirements of such a system was held in January. Two of the main requirements were the need for more robust large shared file-systems and rapid provisioning of hardware and software resources, so that the environment could be scaled up and down in response to changing workloads.

The IT Directorate launched the Standards Working Group (SWG). The purpose of the SWG is to identify "standard products" which provide specific, needed business functionality, with the goal of realizing cost savings by reducing the number of hardware and software products that serve the same function, while at the same time allowing for innovation and the adoption of new technology. The SWG is currently designing procedures for program areas to introduce new products as standards, while at the same time sorting through the current inventory of hardware and software products and deciding what to maintain or recommend for phase out. The group meets weekly. Chad Russell (CSRM) is the primary representative for ADRM, with Shawn Klimek (CES) serving as backup.

After its successful pilot last year, the Data Management Committee (DMC) is continuing to move the development of the Data Management System (DMS) forward. The goal of the DMS is to provide an automated way for users to discover and request data for a particular purpose, and for data owners to review and approve their data for those uses for a particular timeframe. Ultimately, once a project has been approved, the DMS will provision the data and any necessary analytical tools into an environment in which the work can be carried out. Currently, two sub-teams of the DMC are actively working on a communication and training plan, and provisional governance for the initial rollout of the DMS.

The Data Management System (DMS) continues development and is being phased into production. All users with a valid user ID were granted access to the system at the end of June. The goal of the DMS is to provide an automated way for users to discover and request data for a particular purpose, and for data owners to review and approve those uses for a particular timeframe. In the third quarter, the Data Stewardship Executive Policy Committee (DSEP) identified a list of information owners for Census Bureau program areas. The first phase of the DMS rollout will be for information owners to populate the system with information about their datasets and begin relying on the DMS as a request/approval mechanism, with actual delivery of the data occurring outside of the DMS. Data delivery through the DMS will be enabled in the next phase of the rollout, which is expected in late 2012.

Staff: Chad Russell (x33215)

2. RESEARCH

2.1 GENERAL RESEARCH AND SUPPORT (Census Bureau Projects 0351000)

2.2 GENERAL RESEARCH (Census Bureau Projects 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 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 Problem: Compensating for missing data typically involves explicit or implicit modeling. Explicit method includes 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: In FY 2012, staff researched and developed score functions for ranking suspicious records when using selective editing and used these score functions for selective editing of our foreign trade data. We designed measures to effectively compare data edited using selective editing with data edited using ratio edits as implemented in the Edit Master. We used the absolute pseudo-bias and the means and standard errors of the absolute pseudo-bias at different aggregation levels to determine selective editing follow-up cut-off values. Staff also revised previous editing research and computer programs for generating the complete set of ratio edits for a given set of explicit ratio edits.

Staff: María García (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: In FY 2012, in cooperation with staff from the Social, Economic, and Housing Statistics Division (SEHSD), staff developed a new application of the expectation-maximization (EM) algorithm to impute gaps in the wave of data collected by the Survey of Income and Program Participation (SIPP). The application is based on modeling core variables from SIPP leading to the estimation of imputation probabilities for these variables. Then, by conditioning on the imputed values of the core variables, the rest of the variables reported through SIPP can be hot-deck imputed. The results are being documented in a proceeding paper and a *CSRM Research Report*. Staff continued to research the best methods to impute "gaps" in the Survey of Income and Program Participation (SIPP). A first draft of a paper to be presented at the International Statistical Symposium on the Analysis of Longitudinal Data Subject to Measurement Error and Missing Data was completed. The paper presents a model supported by empirical observations from SIPP. The model serves to generate the imputations. The next step: to diagnose-check the model and decide which imputation vehicle - single imputation or multiple imputation - is the most appropriate for the applications the Demographic Directorate is envisioning.

Staff: Yves Thibaudeau (x31706), Chandra Erdman, María García, Martin Klein, Ben Klemens, Rolando Rodriguez

C. Missing Data and Imputation: Multiple Imputation Feasibility Study
Description: Methods for imputing missing data are closely related to methods used for synthesizing sensitive

items for disclosure limitation. One method currently applied to both issues is multiple imputation. Although the two issues may be addressed separately, techniques have been developed that allow data users to analyze data in which both missing data imputation and disclosure limitation synthesis have been accomplished via multiple imputation techniques (e.g., synthetic data). This project ascertains the effectiveness of applying multiple imputation to both missing data and disclosure limitation in the American Community Survey (ACS) group quarters data. Statistical models are used to generate several synthetic data sets for use within the multiple-imputation framework.

Highlights: In FY 2012, further improvements to the TEA package made use of multiple imputation for both missing data and disclosure avoidance feasible for in-house surveys. Work is considered complete on this project.

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

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 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: In FY 2012, staff reviewed the current literature and updated the current list of references on microdata confidentiality. Staff chaired the CSRM talk "Differential Privacy: What It Is, How Staff Hope It Can Be Used, and How You Can Help" by Cynthia Dwork, member of the National Academy of Engineering and head of the privacy group at Microsoft Research, on November 17, 2011. Staff slightly updated the generalized modeling/edit/imputation software. The software can now also used for producing synthetic data with valid analytic properties and reduced/eliminated re-identification risk. Staff continued to circulate information related to privacy and confidentiality to a number of international researchers. Staff served as discussant of three papers on privacy and confidentiality at the FCSM Conference in Washington, DC in January 2012. Staff refereed two papers for Statistical Data Protection 2012.

Due to substantial demand, staff twice taught the 1-day modeling/edit/imputation course that, at the end, covers methods of producing public-use microdata. The basic modeling methods for edit/imputation produce exceptionally high quality imputation models. These models that have been extended with convex constraints (Winkler *Ann. Prob. 1990*) with a very slight reduction in analytic quality and a drastic reduction in re-identification risk. The basic model yields synthetic data in which 80 percent of small cells agree with those in the original microdata (i.e., very high re-identification risk along with very high quality). After applying suitable convex constraints, only 5 percent of the small cells agree with small cells in the original microdata. Staff created methods/software for rapidly analyzing the quality of the models and the data that might be produced using the models. The methods can also be applied in analyzing general classes of loglinear models.

Staff: William Winkler (x34729), William Yancey

B. Noise Multiplication 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 data. Procedures that deliberately alter data prior to their release fall under the general heading of statistical disclosure control. This project develops and studies data analysis under noise perturbation in which data are multiplied by randomly drawn noise variables prior to release. Major goals include (1) developing procedures for drawing inference on population parameters based on noise multiplied data, and (2) comparing these procedures with those based on synthetic data obtained by multiple imputation.

Highlights: In FY 2012, staff developed methods for using noise multiplied data to draw inference on unknown population parameters under an assumed parametric model for the original data. We considered two cases of noise multiplication: (1) all data are confidential and hence are noise perturbed; and (2) only large values above a given threshold are confidential (e.g. income data) and hence only these values are noise perturbed. In each case, we derived general expressions for the likelihood function based on the noise multiplied data. We developed EM algorithms to obtain maximum likelihood estimates of the parameters, and derived expressions for the observed Fisher information to obtain variance estimates and approximate confidence intervals. We derived the details and developed R code for these methods under three specific parametric models: exponential, normal and lognormal. We conducted simulation studies to compare the statistical properties of noise multiplication against those of multiple imputation (or synthetic data) and top coding. We compared these methods through the accuracy of the resulting inferences for population parameters, in terms of bias and mean squared error of the estimates, as well as coverage probability and expected width of confidence intervals. In our simulations, we considered a variety of parameters that may be of interest, such as means, variances, and quantiles. A manuscript describing this work was prepared.

In the work described above, we found that generally if the distribution generating the noise variables has low to moderate variance, then noise multiplied data can yield accurate inferences in several typical parametric models under a formal likelihood based analysis. However, the likelihood based analysis is generally complicated due to the non-standard and often complex nature of the distribution of the noise perturbed sample even when the parent distribution is standard. This complexity places a burden on data users who must either develop the required statistical methods, or implement the methods if already available, or have access to specialized software perhaps yet to be developed. These observations have motivated us to propose an alternate analysis of noise multiplied data based on multiple imputation. Some advantages of this approach are that (1) the data user can analyze the released data as if it were never perturbed in conjunction with simple combining rules; and (2) the distribution of the noise variables does not need to be disclosed to the data user. We have developed these methods for the two cases of noise multiplication mentioned above: (1) each original observation is perturbed and (2) only large values are perturbed. In both cases, we studied two types of imputation procedures as well as three variance estimators that have been proposed in the literature. We performed extensive simulation studies to evaluate the various methods, and to compare the multiple imputation based analysis against a formal likelihood based analysis. We found that the multiple imputation based analysis generally yields only a slight loss of accuracy in comparison to the formal likelihood based analysis. A second manuscript describing this alternative analysis of noise multiplied data was written. Staff presented some aspects of this work in two colloquium talks.

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

C. 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: In FY 2012, staff participated in the 2020 Matching Improvement Process Group headed by DSSD. Staff provided a substantial number of background documents, references lists, and comments to the group. Staff taught the 2-day course "Record Linkage" on January 31-February 1 and again on March 6-7. Each class was full and there is still a waiting list for a possible third teaching of the class. Class participants received lecture notes (300+ pages), extensive lists of references, several background papers, the original *SRD Matching Software*, and name and address standardization software.

Staff provided extensive comments and background papers to members of the 2020 Matching Group, various areas of the Decennial Directorate, and CARRA. Many of the research issues had already been covered in a series of papers from CSRM/SRD and other researchers between 1988 and 2010. The production issues had often been addressed in 20+ matching systems that staff built for person, housing-unit, and business matching in the Decennial and Economic Directorates. The biggest issues were error-rate estimation (Belin and Rubin *JASA 1995*, Winkler 2006, Larsen and Rubin *JASA 2001*, Winkler 2002, Winkler 2004) and 'real-time' matching (as exemplified by the 2010 *BigMatch* Decennial production system written by DSSD and CSRM/SRD). Staff provided very extensive comments about the feasibility of 'real-time' matching because the 2010 *BigMatch* Decennial Census production software was

able to match 10^17 pairs (300 million x 300 million) in 30 hours using 40 CPUs on an SGI Linux box and is also suitable for our very large administrative record tests. The software is on the order of 50 times as fast as commercial and experimental university software and as much as 500 times as fast as software in use at some statistical agencies. Staff provided detailed comments about creating test decks and comparing methods/systems. The comments were based on our extensive experience building systems in 1988-1997 for Decennial Census matching, Agriculture Census matching, Business Register matching, various smaller administrative-records tests by SRD/DSSD, the 1990 Housing Unit Coverage Study, various re-identification, and the enhancement/refinement of the name standardizer and the address standardizer that we give out with the SRD matching software.

Staff (along with staff from DSSD, CARRA, EPCD, and DSMD) met with Director Groves and Deputy Director Mesenbourg to discuss how record linkage research and applications are being coordinated throughout the Census Bureau. Director Groves emphasized the need for having valid methods for adjusting statistical analyses for matching error and methods for estimating record linkage error rates. A number of other groups within the Census Bureau have received our *BigMatch* software (that was used for three production Decennial matching projects in 2010 and will be used for several administrative list projects). All areas of the Census Bureau currently use the name matching software (developed in SRD and the Agriculture Division of the Economic Directorate) and the address matching software (developed by Bill LaPlant of SRD and Brian Beck of the Geography Division). The name and address standardization software still exceed the best commercial software.

Staff wrote and presented the paper "Machine Learning and Record Linkage" that appeared in the *Proceedings of the International Statistical Institute*. Staff also wrote the paper "Cleaning and using administrative lists: Enhanced practices and computational algorithms for record linkage and modeling/editing/imputation" that appeared in the *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Staff wrote and presented the paper "Cleaning and using administrative lists: Methods and fast computational algorithms for record linkage and modeling/editing/imputation." That was a special invited technical paper at the Essnet Conference on Data Integration in Madrid. Staff presented the talk "Record Linkage: Introductory Overview" at the Instituto Nacional de Estadistica in Madrid and an expanded version of the talk with more technical details in the Computer Science Department at the University of Maryland in College Park on February 15, 2012. One of us served as discussant of the DSMD Distinguished Lecture "Factors Affecting Record Linkage in Federal Databases" by Professor Michael Larsen of George Washington University on July 18, 2012. The discussant had earlier reviewed two papers by Professor Larsen and one of his graduate students that related to the talk.

Staff taught the 1-day class on record linkage error-rate estimation. The class covered some of the models and error-rate estimation methods. Staff used the methods in the last three Decennial Censuses, various smaller administrative tests done by SRD/DSSD, and for the systems that staff wrote for the 1992 (and 1997) Agriculture Census list development (12 lists with 16 million records) and for identifying duplicate small businesses in the Business Register. In the latter two situations, the Economic Directorate were able to verify our false-match estimates during field validations or because they had given us test decks for which they knew the true matching status for 20% of the records.

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

D. 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: In FY 2012, staff are still reviewing some of the literature on editing, data quality and imputation, record linkage error-rate estimation and adjusting statistical analyses for record linkage error. Staff updated some of the DISCRETE modeling/edit/imputation software. Ben Klemens and Rolando Rodriguez used an earlier version of the DISCRETE edit/imputation software in the TEA general survey processing software that is being applied in several large Census Bureau projects. Staff put together examples of the high quality of models that can be produced with DISCRETE by relatively junior analysts and programmers. Staff produced methods/software for analyzing the models efficiently and evaluating the effects of the models on improving statistical analyses.

Staff taught the modeling/edit/imputation class twice (48 in first and 43 in second offering) which was a lower-level variant of a short course originally taught in the Institute of Education at the University of London. Students received extensive background material, a few research papers, course notes, and generalized software suitable for production. The software (as with our record linkage software) is portable across all Census Bureau machines. At present, our DISCRETE system (Winkler 1997, 2003, 2008, 2010) is the only system in the world that assures that aggregates of imputed data preserves joint distributions and that individual records satisfy all edits. Further, the methods can scale microdata so that aggregates are closer to external constraints (such as IRS

quintiles for income categories) and create slightly biased models with valid analytic properties with eliminated/reduced re-identification risk (such as is needed for producing public-use files). Our methods and generalized software appear to be the only ones with suitable accuracy and speed for working with national administrative and other files. For the clean-up large files prior to merging, individuals can use DISCRETE and some of the SRD preprocessing/ standardization software. For unduplicating sets of lists, individuals can use *BigMatch* and other matching software from CSRM/SRD. For cleaning sets of merged files used in statistical analyses, and for creating large national files of synthetic data with valid analytic properties and reduced/eliminated re-identification risk, staff can use DISCRETE (although the methods will need enhancement).

Staff gave advice and papers related to methods/software for industry-and-occupation coding (text categorization) and on modeling/edit/imputation to individuals in DSSD and in CSRM. Staff provided advice to DSMD regarding the sampling methods as part of the Demographic redesign. In particular, staff circulated papers/references with explanation of the methods. Staff had earlier reviewed two papers Reid Rottach of DSMD and provided extensive comments on his methods that he has extended for implementation as part of the sample redesign. Staff provided advice, background papers, and references to other staff in CSRM and other parts of the Census Bureau related to machine learning, statistical matching and inverse sampling.

Staff: William Winkler (x34729), William Yancey, María García

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.

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.
• Extend 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.
• Extending small area models to estimators of design-base variance.

A. Small Area Estimation
Description: Methods will be investigated to provide estimates for geographic areas or subpopulations when sample sizes from these domains are inadequate.

Highlights:
Using Small Area Models to Improve Survey Variances:
In FY 2012, staff derived first order Taylor approximations for the replicate weight variance estimator used in the ACS. These approximations give an insight on how to build models which require specifications of means and variances of the survey variance estimators. Prior specification of the variance of the survey variance estimator was determined by empirical results. This current approach leads to a more theory driven model. We found through simulation studies, that the first order approximations worked well for estimating the design-based variance, even for small sample sizes and this suggests a form of a generalized variance function (GVF) that can be used for estimates based on small sample sizes. The first-order approximation overestimated the relative variance of the variance estimator and further work is ongoing to understand and fix this problem.

Diagnostics of Model Misspecification in Small Area Models:
Staff plans to evaluate small area models more area-level data when the models are potentially misspecified. In the Fay-Herriot model the normality assumption of the linking model may not hold. Staff plans to investigate the Mean Square Estimation approximation based on the

maximum likelihood estimator of the parameters. The goal is to develop diagnostics for misspecification using robust sandwich-formula variances, cross-validation, and others. This project is currently in the planning stage.

Bayesian Inference Using a Design-Adjusted Likelihood: Statistical model building often starts from the assumption that data are independent or obtained from a simple random sample. However, using data from a survey often introduces layers of selection, weighting and clustering, and the problem of accounting for the design through a model may necessitate extremely large models with many parameters. In FY 2012, staff worked on the problem of approximating the likelihood with design-adjusted models which are simpler to work with through use of the design effect, and compared the approximation to the exact likelihood in special cases for small samples. This problem is particularly important for Bayesian small area modeling which requires specification of a likelihood and a prior distribution.

ACS Simulation Study Software:
Staff built a software tool for empirical evaluation and comparison of small-area models against data from an ACS-like population. The tool creates a pseudo-population from unit-level ACS data, and then approximates the ACS's complex sampling and weighting process to draw samples repeatedly. Small area models can be fitted on each of many such samples, allowing researchers to evaluate each model's fit and sampling variability. Staff completed an initial version of the tool and used it to evaluate models for the JSM proceedings paper "A Bayesian Zero-One Inflated Beta Model for Small Area Shrinkage Estimation" (Wieczorek, Nugent, and Hawala, 2012). Staff gathered user requirements and advice from other researchers (Maples, Bell, Datta, and Franco) to ensure that the next version is more sophisticated, more broadly applicable, and easier to use.

Staff: Jerry Maples (x32873), Aaron Gilary, Ryan Janicki, Jerzy Weiczorek, Gauri Datta

B. 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.

Highlights: In FY 2012, this project was initiated with discussions and preliminary drafts between staff. *S*taff continued previous discussion and preliminary write-ups to develop this idea, by considering alternative small area models which might hold on subsets of the set of small areas.

Staff: Eric Slud (x34991), Gauri Datta

C. Visualization of Small Area Estimates
Description: Methods are needed to display estimates for a large number of small areas or domains. Displays should accurately convey the level of statistical uncertainty and should guide readers in making comparisons appropriately between domains or over time.

Highlights: In FY 2012, Staff gave presentations both inside and outside of the Census Bureau on using an in-house interactive visualization to view small area estimates and their uncertainties. In collaboration with visiting scholars from Cornell, University of Tennessee, and the University of Colorado, staff created novel visualizations (both static and animated) of statistical uncertainty in maps of ACS data. Staff also taught a workshop on creating graphics with the statistical software R to the "Dev3" Data Visualization IOE subteam members.

Staff: Eric Slud (x34991), Gauri Datta

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, (1) estimates of low-valued exports and imports not currently reported, (2) influential values in retail trade survey, and (3) 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 10 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 thorough 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. Survey Productivity and Cost Analysis
Description: The Survey Productivity and Cost Analysis (SPCA) Group has been established as a cross-directorate analytic team to conduct methodological research toward the goal of continuous improvement in survey operational efficiency. The group will both initiate and respond to issues related to survey performance indicators including cost, data quality, and data collection progress, as they relate to survey design. Our Center is represented on this team along with staff from the Research and Methodology Directorate, the Demographic Programs Directorate, the Decennial Directorate, the Center for Economic Studies (CES), the Field Directorate, and the Center for Survey Measurement (CSM).

Highlights: In FY 2012, in partial fulfillment of the 2010 Operational Efficiency Award, the SPCA staff developed numerous charts and a dashboard for monitoring survey productivity and cost. These tools are now being used by field staff across the regions and are documented in the internal report "Exploring the Use of Paradata and Cost Data for Survey Management," prepared by Matthew Jans (CSM) and Barbara O'Hare (DIR), with content contributed by Chandra Erdman (CSRM), Tamara Adams (DSSD), Cathy Buffington (CES), and David Morgan (FLD) (2012).

The team also developed discrete-time hazard models for estimating response propensities in the Current Population Survey, the American Community Survey, the Survey of Income and Program Participation, the National Crime Victimization Survey, the National Health Interview Survey, and the Consumer Expenditure Survey. These models were presented at the American Association for Public Opinion Research 67th Annual Conference. Finally, the SPCA team developed new methodology for setting response rate standards in

surveys conducted by the Census Bureau and presented this work at the 23rd International Workshop on Survey Nonresponse, held at Statistics Canada in Ottawa, Canada.

Staff: Chandra Erdman (x31235), Julie Tsay

B. Household Survey Design and Estimation
[See Project 5385260, Decennial Directorate – American Community Survey (ACS)]

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: In FY 2012, collaborating with a team in the Economic Directorate, staff conducted research to find an automated method for detecting and treating verified influential values in the Monthly Retail Trade Survey (MRTS) to replace the current, more subjective procedure performed by analysts. The results led to the decision to pursue a weighted M-estimation as the methodology to test on a side-by-side basis real time with the MRTS. Work has begun on methodology for tailoring the parameter settings for each industry in MRTS and the Monthly Wholesale Trade Survey. The team used a simulation methodology to investigate the relative bias and RRMSE for two candidate treatments, M-estimation and Clark Winsorization, under several realistic scenarios for influential values. The simulation found both are effective but Clark Winsorization trims about 0.5 percent of the observations when no influential value is present, introducing adjustments that achieve a very small reduction in MSE although the RRMSE for the two methods are approximately equal. The trimming has a practical disadvantage in the MRTS application because the staff usually researches the accuracy of each observation that is flagged as influential. Investigating minor changes is an inefficient use of staff time. The research identified one parameter, the initial value of the tuning constant ϕ, which affects the size of the M-estimation detection region. The ability to set the initial ϕ permits some control of the amount of staff time that has to be devoted to checking. Presentations on the results were given at the 2012 International Conference on Establishment Surveys and the 2012 Joint Statistical Meetings and papers were submitted to the proceedings for each conference.

Staff also worked with a team in Governments Division to investigate methodology for detecting and treating influential values in the Annual Survey of Government Employees. A simulation study investigated the performance of Clark Winsorization and a variation on the application of M-Estimation. Based on the performance measures of relative bias and RRMSE of the estimate of total employment, M-estimation was superior to Clark Winsorization. This work was presented at the Federal Committee on Statistical Methodology Research Conference and a paper was submitted to the conference proceedings.

Staff: Mary Mulry (x31759)

D. Ranking Methodology Development and Evaluation
Description: This project undertakes research into the development and evaluation of statistical procedures for using sampled 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 quantities of interest.

Highlights: In FY 2012, staff systematically reviewed several nonparametric statistical methods based on ranks as presented in the book, *Nonparametric Statistical Methods, 2nd Edition*, by Myles Hollander and Douglas A. Wolfe. We have also reviewed the application of parametric bootstrap methods, and some graphical procedures for quantifying and conveying the amount of uncertainty in an estimated ranking. We reviewed selected related published papers. We initiated a systematic review of the bootstrap; our coverage has included basic theory for nonparametric and parametric bootstrap estimation of variance, bias, and sampling distribution. We presented these topics in a series of lecture style meetings, and implemented some of the methods in R. We applied some methods to public use data from the American Community Survey. New test statistics and their sampling distribution were also explored.

Staff: Tommy Wright (x31702), Martin Klein, Jerzy Wieczorek, Derrick Simmons, John Zylstra, Joel Beard, Charles Davis, Barrett Veazey

E. Specification of ACS Estimands Under Multiple Housing Unit (HU) Weighting Adjustments

Description: The ACS HU weighting-adjustment process is constructed in numerous stages, with successive controls and raking adjustments applied in a specific order to higher level aggregates in geography and time. The objective of this project is to understand exactly what characteristic of the frame population is being estimated by this process for each survey attribute, both under ideal conditions (related to geographic and temporal homogeneity) and under actual conditions. The activity will involve modeling and, ultimately, simulation to understand the effects of current and alternative weighting adjustment methods.

Highlights: In FY 2102, staff reviewed ACS documentation and discussed formal models which might capture the effects of successive ACS weighting steps, and prepared and circulated a preprint summarizing these. Further comments on and revisions of this draft are expected to result in data-analytic and simulation research in the next fiscal year based on ACS data.

Staff: Eric Slud (x34991), Lynn Weidmann, Tucker McElroy, Michael Ikeda, Patrick Joyce, Martin Klein

F. Respondent-Driven Sampling for Hidden and Hard-to-Reach Populations

Description: This project investigates a statistical application of social network analysis called respondent-driven sampling (RDS). RDS is a form of snowball sampling which allows researchers to make estimates about hidden or hard-to-reach populations. An RDS sample of a target population is collected by selecting "seed" respondents who then (with financial incentive) recruit their friends to be future respondents. Because respondents are not selected by simple random sampling, care must be taken when making estimates from this type of sample. It has been shown that if certain conditions are met and if the appropriate estimation procedures are used, then estimates from an RDS sample are unbiased. This project will further examine the ability of RDS to produce unbiased estimates, as well as the effect of real-world violations of key assumptions on the model.

Highlights: In FY 2012, staff investigated the feasibility of social network sampling methods for use in Census Bureau programs. Dr. Matthew Salganik, Assistant Professor of Sociology at Princeton University, was invited for *SUMMER AT CENSUS* and presented two seminars, entitled "Assessing Respondent-Driven Sampling," and "Generalized Network Scale-Up Method for Estimating the Sizes of Hard-to-Count Groups: Evidence from Brazil and Rwanda."

As a result of these seminars and discussions with Salganik, it was concluded that respondent-driven sampling involves a variety of limitations that may hinder immediate applications, but future areas of research were discussed. Additionally, the generalized scale-up method has been shown to hold more promise in potential applications, which will continue to be explored.

Staff: Taniecea Arceneaux (x33440)

G. 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 2012, staff linked these two problems that have well-known but different solutions; one solution is not explicitly exact (Neyman allocation), and the other (equal proportions) is exact. We gave explicit, exact solutions for both and noted that the solutions are equivalent. In fact, we conclude by showing that both problems are special cases of a general problem. The result is significant for stratified random sampling in that it explicitly shows how to minimize sampling error while keeping the overall sample size fixed at n. An example reveals that controlled rounding with Neyman allocation does not always lead to the optimum allocation (Wright, 2012).

Staff: Tommy Wright (x31702)

Statistical Computing and Software

Motivation: Modern statistics and computing go hand in hand, and new statistical methods need to be implemented in software to be broadly adopted. The focus of this research area is to develop general purpose software using sound statistical methods that can be used in a variety of Census Bureau applications.

These application areas include: survey processing - editing, imputation, non-response adjustment, calibration and estimation; record linkage; disclosure methods; time series and seasonal adjustment; variance estimation; small-area estimation; and data visualization, exploratory data analysis and graphics. Also see the other sections in this document for more detail on some of the topics.

Research Problems:

- Investigate the current best and new statistical methods for each application.
- Investigate alternative algorithms for statistical methods.
- Determine how to best implement the statistical algorithms in software.

Potential Applications:
• Anywhere in the Census Bureau where statistical software is used.

A. TEA Software Development
Description: 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, we can guarantee that all imputations can pass edit requirements, and we 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: In FY 2012, staff made improvements to TEA to improve support of island area group quarters disclosure avoidance. Staff began developing an internal and external distribution method for TEA, in collaboration with staff in the Center for Applied Technology. Staff added additional features to the decision making model of respondents, using a typology of mindsets from Bates, Mulry, et al. The propensity to respond for each mindset was calibrated to produce an optimal fit between the model predictions and 2000 NRFU data. [See also Decennial Project 6510201(A).]

Staff: Ben Klemens (x36864)

B. R Users Group
Description: The initial objective of the R User 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: In FY 2012, staff organized three meetings for R users. The group of users has increased to around 60 members with varied interests related to the R software. The first meeting, which took place February 21, had the intent to elicit the interests of the participants. Four topics were identified: introductory R, edit and survey processing, modeling and simulation, and graphics. The second and third R-users meetings were intended to meet the interests of those users interested in introductory R information. In that context, Jerzy Wieczorek (SEHSD, now in CSRM) gave two introductory lectures to the group.

Staff met with the Census Bureau IOE staff to establish a long-term strategy to safeguard and support the use of R in research and production projects at the Census Bureau. Decisions were made to initiate discussions with the IT Directorate on how to best integrate R in the new IT architecture at the Census Bureau. Specifics include the location of an R server and security protocols for downloading/upgrading R packages.

Staff: Yves Thibaudeau (x31706)

C. Web Scraping Feasibility Investigation
Description: The goal of this project is to investigate the feasibility of developing and implementing a Web scraping tool. This tool will collect publicly available information posted by businesses. Knowledge of this auxiliary information may be useful in improving estimates with economic data.

Highlights: In FY 2012, staff designed a Java program to extract data from .xml files. The advantage of having this functionality is that the data that is piped into an .xml file from either Python scrapy (the web scraping program written last year) or another web scraping toolkit that is now easy to download directly into iMetrica for easy data control, plotting, signal extraction, etc. The program should be used in conjunction with the Python code for web scraping which was written last year.

Staff: Chris Blakely (x31722)

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-12-ARIMA 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-12-ARIMA 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.

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

Highlights: During FY 2012, staff: (a) developed the primary alias filter discretization method for converting continuous-time signal extraction filters into discrete filters appropriate for stock or flow time series. Code was developed to compute filter coefficients and assess the discretization mean squared error, and numerical results were computed; (b) made minor edits and further empirical work to paper handling signal extraction for co-integrated vector time series. The distinction between co-integrated and co-linear time series was elucidated, so that the scope of the theoretical results is clear; (c) resolved the issue of identifiability for a class of hysteric structural models appropriate for cases where underlying stochastic components are cross-correlated. The empirical results were also modified; (d) continued developments on a survey paper discussing the Direct Filter Approach of Marc Wildi. Staff sponsored a visit by Wildi to facilitate research. Results for the nonstationary case were added, and numerical results for the Hodrick-Prescott filter were encoded and computed; (e) implemented a multi-step ahead forecasting fitting function based on finite-sample predictors, and made other revisions needed for the paper on this topic; (f) continued theoretical work on spectral density estimation theory using fixed bandwidth ratio asymptotics, with a development of vanishing bandwidth as well. The goal here is to take the Laplace Transform of Gaussian quadratic forms, and derive a practical expansion of the cumulative distribution function in terms of known basis functions; (g) investigated fitting time series models by a criterion that seeks to minimize signal extraction revision variance. Method was encoded and begun testing on retail series.

Staff: Tucker McElroy (x33227)

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

Highlights: During FY 2012, staff did the following: (a) continued empirical work on the study of HAC estimators under the context of long memory or negative memory time series. Staff also derived further functional limit theorems for a generalization to nonparametric spectral density estimates in the same stochastic context; (b) continued theoretical work on the sample autocovariances and autocorrelations for linear heavy-tailed long memory time series, and generalized the consistency of the subsampling distribution estimator to the case of higher memory by imposing a smaller block size; (c) derived new results on the sample autocovariances and autocorrelations for GARCH processes, and derived effective self-normalizations for these statistics when the underlying data is heavy-tailed; (d) derived new formulas comparing direct and iterative forecasting from finite samples, and encoded algorithms to fit multi-step ahead forecasting models. Staff made extensive numerical comparisons of resulting mean squared error for competing methods (poster presented at OxMetrics conference at George Washington University); (e) developed generic forecasting formulas for non-stationary vector time series that might be co-integrated, including both semi-infinite and finite samples, along with mean squared error formulas. Method has been encoded and empirical work completed, with applications to fertility rates and housing starts; (f) continued refinements to theoretical work on estimation of constrained vector time series models using the Whittle likelihood, maximum likelihood estimation, and the method of moments. The stability parameter manifold was also explored by the staff, with a natural spherical parametrization that allows for parameter constraints; (g) continued making improvements to algorithm for the vector exponential time series model. Because commutativity is not preserved by the matrix exponential, a revised formulation of the Wold decomposition was necessary. An efficient algorithm that bypasses an exponentially growing amount of combinatorics was discovered, and Bayesian implementation is complete; (h) continued work on the study of popular tail index estimators from the perspective of taking the number of upper order statistics to be a fixed fraction of sample size. The asymptotic expressions for bias and variance have now been confirmed by preliminary empirical work, and optimal tuning parameters have been studied, along with a method to reduce bias via regression; (i) developed efficient algorithm to compute autocovariances of a process with multiple long range dependent poles and/or zeroes in its spectral density. Staff tested this method on weekly and monthly seasonal data as well as sunspot data; (j) further developed the testing framework for the Generalized Likelihood Ratio statistic when the ARMA

models are non-nested, with specific examples developed as well.

Staff: Tucker McElroy (x33227), David Findley

C. Spatial Statistics Study Group
Description: This group meets to keep up to date with and to discuss published research on spatial statistics.

Highlights: In FY 2012, the study group met weekly from January through May, reading one chapter of a reference book at a time, presented by a different participant each week. Staff also completed theoretical and algorithmic work on a cepstral random field model for lattice spatial data, giving very general asymptotic results (both frequentist and Bayesian) for estimating model parameters with regression effects.

Staff: Tucker McElroy (x33227)

Experimentation, Simulation, and Modeling

Motivation: Experiments at the Census Bureau are used to answer many research questions, especially those related to testing, evaluating, and advancing survey 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. Monte Carlo simulation techniques aid in the design of complicated experiments as well as the evaluation of complex statistical models.

Research Problems:
• 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).
• Develop methods for designed experiments embedded in sample surveys. Simulation studies can provide further insight (as well as validate) any proposed methods.
• 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 ACS.

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: In FY 2012, staff gathered data relating response propensities to demographics and geography to calibrate the existing model. Staff also worked with CARRA (Center for Administrative Records Research & Applications) to determine project focus and long term goals. To accommodate its use for the Island Areas, ACS staff added a missing data model to TEA. Although it is primarily used to edit and impute individual-level characteristics, staff researched methods for adapting TEA for household-level counts, which would broaden the package's application to situations such as the Decennial Census.

Staff: Ben Klemens (x36864)

B. Tolerance Intervals for the Difference of Proportions
Description: This project undertakes the investigation of tolerance intervals for the difference of two sample proportions. Unlike a confidence interval, which provides bounds for the difference between population proportions, a tolerance interval provides statistical bounds within which a certain proportion of the difference between the sample proportions is expected to lie, with a specified confidence level. As an example, suppose it is desired to estimate the difference in the proportion of households who do not have health coverage among two ethnic groups. Since the estimated difference between the proportions is subject to sampling variability, having bounds on the estimated difference clearly provides valuable information. In this research, tolerance intervals will be developed when sampling with or without replacement. Coverage studies of the

proposed intervals will also be performed. While the normal approximation can be used for the computation of such tolerance intervals, such an approximation will be unsatisfactory when the sample sizes are not large and/or the proportions are small. The small sample scenario is relevant in social surveys where there is need to estimate proportions and difference between proportions in many small groups or small domains. This research will focus on the development of accurate tolerance intervals in a small sample setting.

Highlights: In FY 2012, staff tested a fiducial-based approach to handle the tolerance interval computation. Computational issues have also been addressed. Extensive simulation studies showed that the coverage probabilities are near nominal for most settings. When either (or both) of the groups have a small sample size (e.g., less than 10), the coverage probabilities tend to be conservative. However, we investigated a bootstrap calibration to provide an adjustment for these small sample size cases.

We applied our methodology to data from the 2010 U.S. Census Nonresponse Follow-Up. The work is complete and a manuscript has been submitted for publication.

Staff: Thomas Mathew (x35337), Derek Young

C. An Inverse Sampling Plan
Description: When conducting an experiment, fixing sample sizes may result in too few observations relative to the conditions of the experiment. To overcome this difficulty, one may adopt an inverse sampling design that fixes the number of events, which results in random variables following a negative binomial distribution. For planning purposes and cost evaluations, various projects conducting test surveys (e.g., ACS) could benefit from projecting statistical bounds on the number of expected nonresponses (and thus the total sample size) using an inverse sampling plan based on previous data from similar surveys. Such bounds are essentially tolerance limits for negative binomial random variables.

Highlights: In FY 2012, staff identified a methodology for constructing tolerance intervals for negative binomial random variables. Since this approach relies on constructing confidence intervals for the negative binomial proportion, eight different methods were assessed.

Coverage studies were performed in R to identify which of these eight approaches yields the best coverage probabilities under different conditions. A manuscript was prepared, submitted, and accepted to a journal.

Staff: Derek Young (x36347)

Research and Development Contracts

A. Research and Development Contracts
Description: The Research and Development Contracts are indefinite delivery, indefinite quantity task order contracts for the purpose of obtaining contractor services in highly technical areas to support research and development activities across all Census Bureau programs. The contracts provide a pool of contractors to assist the Census Bureau in conducting research on all survey and census methods and processes to improve our products and services. The prime contractors include educational institutions, university supported firms, and privately owned firms that concentrate in sample survey research, methodology, and applications to create a pool of specialists/experts to tackle some of the Census Bureau's most difficult problems through research. Many of the prime contractors are teamed with one or more organizations and/or have arrangement with outside experts/consultants to broaden their ability to meet all of the potential needs of the Census Bureau. These five-year contracts allow Census Bureau divisions and offices to obtain outside advisory and assistance services to support their research and development efforts quickly and easily.

<u>R&D 2007 Contracts</u>
During FY 2012, eight (8) modifications were awarded and three (3) task orders were completed. To date, there have been ninety-six (96) task orders awarded under the R&D 2007 contracts, with a monetary value of over $126 million dollars. Ninety-one (91) task orders have been completed and one task order terminated, leaving four (4) active tasks.

<u>R&D 2014 Contracts</u>
During FY 2012, twelve (12) new task orders were awarded, forty-seven (47) modifications were awarded and fifteen (15) task orders were completed. To date, there have been fifty-nine (59) task orders awarded under the R&D 2014 contracts with a monetary value of over $46 million. Thirty-four (34) task orders have been completed, leaving twenty-five (25) active task orders.

Staff: Ann Dimler (x34996), Esan Sumner

Summer at Census

Description: Recognized scholars in the following and related fields applicable to censuses and large-scale sample surveys are invited for short-term visits (one to ten 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: Sponsored, with divisions around the Census Bureau, scholarly, short-term visits by 26 researchers/ leaders who collaborated extensively with us and presented seminars on their research. For a list of the 2012 *SUMMER AT CENSUS* scholars, see *http://www.census.gov/research/summer_at_census/ summer_2012.php.*

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, Gloria Prout, Esan Sumner, Kelly Taylor

3. PUBLICATIONS

3.1 JOURNAL ARTICLES, PUBLICATIONS

Alexandrov, T., Bianconcini, S., Dagum, E., Maass, P., and McElroy, T. (2012). "The Review of Some Modern Approaches to the Problem of Trend Extraction," *Econometric Reviews,31*, 593-624.

Bates, N. and Mulry, M. (2011). "Using a Geographic Segmentation to Understand, Predict, and Plan for Census and Survey Mail Nonresponse," *Journal of Official Statistics, 27(4),* 601–618.

Chen, Q., Elliot, M.R., and Little, R.J. (In Press). "Bayesian Inference for Finite Population Quantiles from Unequal Probability Samples," *Survey Methodology.*

Findley, D. and Quenneville, B. (In Press). "Uncorrelatedness and Other Correlation Options for Differenced Seasonal Decomposition Components of ARIMA Model Decompositions," *Taiwan Economic Forecast and Policy*.

Findley, D. and Quenneville, B. (In Press). "The Timing and Magnitude Relationships Between Month-to-Month Change and Year-to-Year Changes That Make Comparing Them Difficult," *Taiwan Economic Forecast and Policy.*

Findley, D., Monsell, B., and Hou, C.-T. (In Press). "Stock Series Holiday Regressors Generated from Flow Series Holiday Regressors," *Taiwan Economic Forecast and Policy.*

Hunter, D. and Young, D. (2012). "Semiparametric Mixtures of Regressions," *Journal of Nonparametric Statistics, 24(1),* 19-38.

Jach, A., McElroy, T., and Politis, D. (2012). "Subsampling Inference for the Mean of Heavy-tailed Long Memory Time Series," *Journal of Time Series Analysis, 33,* 96-111.

Janicki, R. and Malec, D. (In Press). "A Bayesian Model Averaging Approach to Analyzing Categorical Data with Nonignorable Nonresponse," *Computational Statistics and Data Analysis.*

Klein, M., Neerchal, N., Sinha, B., Chiu, W., and White, P. (In Press). "Statistical Inferences From Serially Correlated Methylene Chloride Data," *Sankhya B.*

Little, R.J. (In Press). "Calibrated Bayes: An Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder)," *Journal of Official Statistics.*

Little, R.J., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., Neaton, J.D., Shih, W., Seigel, J.P., and Stern, H. (2012). "The Design and Conduct of Clinincal Trials to Limit Missing Data," *Statistics in Medicine,* early views. DOI: 10.1002/sim.5519.

Little, R.J., D'Agostino, R., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., Frangakis, C., Hogan, J.W., Molenberghs, G., Murphy, S.A., Rotnitsky, A., Scharfstein, D., Neaton, J.D., Shih, W., Seigel, J.P, and Stern, H. (In Press). "The Prevention and Treatment of Missing Data in Clinical Trials," *New England Journal of Medicine.*

McElroy, T. (2012). "An Alternative Model-based Seasonal Adjustment that Reduces Over-Adjustment," *Taiwan Economic Policy and Forecast, 43,* 35-73.

McElroy, T. (In Press). "Forecasting CARIMA Processes with Applications to Signal Extraction," *Annals of the Institute of Statistical Mathematics.*

McElroy, T. (2012). "The Perils of Inferring Serial Dependence from Sample Autocorrelation of Moving Average Series," *Statistics and Probability Letters, 82,* 1632-1636.

McElroy, T. and Holan, S. (In Press). "A Conversation with David Findley," *Statistical Science.*

McElroy, T. and Holan, S. (2012). "On the Estimation of Autocovariances for Generalized Gegenbauer Processes," *Statistica Sinica, 22,* 1661-1687.

McElroy, T. (2011). "A Nonparametric Method for Asymmetrically Extending Signal Extraction Filters," *Journal of Forecasting, 30,* 597-621.

Durrant, M., McElroy, T., and Durrant, L. (2012). "First Metatarsophalangeal Joint Motion in Homo Sapiens: Theoretical Asociation of Two-Axis Kinematics and Specific Morphometrics," *Journal of the American Podiatric Medical Association ,102,* 374-389.

McElroy, T. and Politis, D. (2012). "Fixed-b Asymptotics for the Studentized Mean for Long and Negative Memory Time Series," *Econometric Theory, 28,* 471-481.

Shao, J., Klein, M., and Xu, J. (In Press). "Imputation for Nonmonotone Nonresponse in the Survey of Industrial Research and Development," *Survey Methodology.*

Slud, E. (2011) book review, "Essential Methods for Design Based Sample Surveys," *Journal of the American Statistical Association,* to appear (available online 3/20/2012).

Slud, E. (2012), "Assessment of Zeroes in Survey-Estimated Tables via Small-Area Confidence Bounds," *Journal of the Indian Society for Agricultural Statistics,* Special Issue on Small Area Estimation, eds. R.Chambers, U.C. Sud and H. Chandra, v.66, pp. 157-169.

Young, D. (2012), book review of "Optimal Experimental Design with R," by D. Rasch, J. Pilz, R. Verdooren, and A. Gebhardt." *Journal of Applied Statistics, 39(8),* 1848-1849.

Young, D. (In Press). "Regression Tolerance Intervals," *Communications in Statistics - Simulation and Computation.*

Young, D. (In Press). "A Procedure for Approximate Negative Binomial Tolerance Intervals," *Journal of Statistical Computation and Simulation.*

West, B. and Little, R.J. (In Press). "Nonresponse Adjustment Based on Auxiliary Variables Subject to Error," *Applied Statistics.*

Wright, T. (2012). "The Equivalence of Neyman Optimum Allocation for Sampling and Equal Proportions for Apportioning the U.S. House of Representatives," *The American Statistician, 66 (4),* 217-224.

3.2 BOOKS/BOOK CHAPTERS

Bell, W., Holan, S., and McElroy, T. (Editors). (2012). *Economic Time Series: Modeling and Seasonality*. New York: Chapman Hall.

Holan, S. and McElroy, T. (2012). "On the Seasonal Adjustment of Long Memory Time Series." In W. Bell, S. Holan, and T.McElroy (Eds.), *Economic Time Series: Modeling and Seasonality.* New York: Chapman and Hall.

McElroy, T. and Holan, S. (2012). "The Error in Business Cycle Estimates Obtained from Seasonally Adjusted Data." In W. Bell, S. Holan, and T. McElroy (Eds.), *Economic Time Series: Modeling and Seasonality*. New York: Chapman and Hall.

Mulry, Mary H. (2011). "Post-enumeration Survey." In M. Anderson, C. Citro, and J. Salvo. (Eds.), *Encyclopedia of U.S. Census, 2nd Edition: From the Constitution to the American Community Survey*. Washington, DC.: CQ Press, 339 – 343.

Schafer, Joseph. (In Press). "Bayesian Penalized Spline Models for Statistical Process Monitoring of Survey Paradata Quality Indicators." In Frauke Kreuter (Ed.), *Improving Surveys with Paradata: Analytic use of Process Information*. Hoboken, NJ: Wiley.

Winkler, W. E.. (In Press). "Record Linkage," in *Encyclopedia of Environmetrics*, New York: John Wiley & Sons.

Wright, T. (In Press). "U.S. Bureau of the Census," in *Handbook of Behavioral and Social Sciences*, New York: John Wiley & Sons. (Article is slightly updated version of an article by the same title that appeared in the *Encyclopedia of Statistical Sciences* (2006), John Wiley & Sons, Inc.)

3.3 PROCEEDINGS PAPERS

World Congress of Statistics, International Statistical Institute, Dublin Ireland, August 21 – August, 26, 2011.

- Winkler, W. E., "Machine Learning and Record Linkage."

Joint Statistical Meetings, American Statistical Association , Miami, FL, July 31 – August 3, 2011.
2011 Proceedings of the American Statistical Association

- William Winkler, "Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Editing/Imputation," 108-117.
- Mary Mulry and Bruce Spencer, "Designing Estimators of Nonsampling Errors in Estimates of Components of Census Coverage Error," 5309-5323.
- Aaron Gilary, "Recursive Partitioning for Racial Classification Cells," 2706-2720.
- Ryan Janicki, "Selection of Prior Distributions for Multivariate Small-Area Models with Applications to Small-Area Health Insurance Estimates," 4170-4184.
- Eric Slud, J. Suntornchost, and R. Wei, "An Age-segmented Poisson Log-Bilinear Model for Sex- and Cause-Specific Mortality Rates," 1801-1815.

UN/ECE Work Session on Statistical Data Editing, Oslo, Norway, September 24 – 26, 2012.

- Maria Garcia, "An Application of Selective Editing to the US Census Bureau Trade Data." http://www.unece.org/stats/documents/2012.09.sde.html

3.4 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY RESEARCH REPORTS

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

RR (Statistics #2012-01): Michael Ikeda, Julie Tsay, and Lynn Weidman. "Exploratory Analysis of the Differences in American Community Survey Respondent Characteristics between the Mandatory and Voluntary Response Methods," January 11, 2012. (Also 2012 American Community Survey Research and Evaluation Report Memorandum Series #ACS12-RER-02.)

RR (Statistics #2012-02): Patrick M. Joyce, Donald Malec, Roderick A. Little, and Aaron Gilary. "Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations," January 24, 2012.

RR (Statistics #2012-03): Michael Beaghen, Tucker McElroy, Lynn Weidman, Mark Asiala, and Alfredo Navarro. "Interpretation and Use of American Community Survey Multiyear Estimates," April 18, 2012.

RR (Statistics #2012-04): Elizabeth T. Huang and William R. Bell. "An Empirical Study on Using Previous American Community Survey Data Versus Census 2000 Data in SAIPE Models for Poverty Estimates," April 25, 2012.

RR (Statistics #2012-05): Benoit Quenneville and David F. Findley. "The Timing and Magnitude Relationships Between Month-to-Month Changes and Year-to-Year Changes That Make Comparing Them Difficult," May 9, 2012.

RR (Statistics #2012-06): David F. Findley. "Uncorrelatedness and Other Correlation Options for Differenced Seasonal Decomposition Components of ARIMA Model Decompositions," May 29, 2012.

RR (Statistics #2012-07): Roderick J. A. Little. "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics," July 9, 2012.

RR (Statistics #2012-08): Tucker S. McElroy and Agustin Maravall. "Optimal Signal Extraction with Correlated Components," July 11, 2012.

RR (Statistics #2012-09): Tucker S. McElroy and Brian C. Monsell. "Model Estimation, Prediction, and Signal Extraction for Nonstationary Stock and Flow Time Series Observed at Mixed Frequencies," July 11, 2012.

RR (Statistics #2012-10): Maria Garcia and Emily Bartha. "Score Functions for Selective Editing of the US Census Bureau Trade Data," August 20, 2012.

RR (Statistics #2012-11): Tucker McElroy and Marc Wilidi. "Multi-Step Ahead Estimation of Time Series Models," September 11, 2012.

RR (Statistics #2012-12): William R. Bell, Gauri S. Datta, and Malay Ghosh. "Benchmarking Small Area Estimators," September 12, 2012.

RR (Statistics #2012-13): Aditi Ramachandran, Lisa Singh, Edward H. Porter, and Frank Nagle. "Exploring Re-identification Risks in Public Domains," September 21, 2012.

RR (Statistics #2012-14): Tucker S. McElroy and Brian C. Monsell. "The Multiple Testing Problem for Box-Pierce Statistics," September 24, 2012.

RR (Statistics #2012-15): Ryan Janicki and Eric V. Slud. "Effects of Missing Data on Modeling Enumeration Status in the U.S. Census," September 24, 2012.

3.5 OTHER REPORTS

Sinsheimer, J.S., Little, R.J., and Lake, J.A., (2012). "Rooting Gene Trees without Outgroups: EP Rooting," *Genome Biology and Evolution*. DOI: 10.1093/gbe/evs/047.

4. TALKS AND PRESENTATIONS

University of Maryland, College Park, MD, October 6, 2011.
- McElroy, Tucker, "Signal Extraction for Nonstationary Multivariate Time Series."

Conference on Biometric Risk Assessment, Silver Spring, MD, October 14, 2011.
- Slud, Eric, "Parametric Survival Densities from Phase-Type Models."

National Bureau of Statistics of China, Beijing, People's Republic of China, October 24-27, 2011.
- Monsell, Brian, "Seasonal Adjustment Class."

2011 Marketing Outlook Forum, U.S. Travel Association, Fort Worth, TX, October 26-27, 2011.
- Mulry, Mary H., "Population Changes through the Lens of the 2010 Census & More."

Statistics Canada International Methodology Symposium, Ottawa, Ontario, Canada. November 1-4, 2011.
- McElroy, Tucker, "The Error in Business Cycle Estimates Obtained from Seasonally Adjusted Data."

Mathematical Association of America, Florida Chapter Meeting, University of West Florida, November 18, 2011.
- Klein, Martin, "Statistical Analysis based on Physiologically-based Pharmacokinetic Models."

Instituto Nacional de Estadistica, Madrid, Spain, November 23, 2011.
- Winkler, William E., "Record Linkage: Introductory Overview."

Essnet Data Integration Conference, Madrid, Spain, November 24, 2011.
- Winkler, William E., "Cleaning and Using Administrative lists: Methods and Fast Computational Algorithms for Record Linkage and Modeling/Editing/Imputation."

Seminar, The George Washington University, Washington, D.C., December 8, 2011.
- McElroy, Tucker, "When are Direct Multi-Step and Iterative Forecasts Identical?"

Federal Committee on Statistical Methodology Conference, Washington, DC, January10-12, 2012.
- Winkler, William E., "Discussion of Three Papers on Privacy and Confidentiality."
- Barth, Joseph James, John Tillinghast, and Mary H. Mulry, "Influential Values and Robust Estimation in the Annual Survey of Public Employment and Payroll."
- Maples, Jerry, "Discussion on Small Area Estimation Session."

Department of Computer Science, University of Maryland, College Park, MD, February 15, 2012.
- Winkler, William E., "Record Linkage: Introductory Overview."

Bureau of Labor Statistics, Washington, D.C., March 1, 2012.
- Monsell, Brian, "Update on the Development of X-13ARIMA-SEATS."

OxMetrics Conference, The George Washington University, Washington, D.C., March 16, 2012.
- McElroy, Tucker, "When are Direct Multi-Step and Iterative Forecasts Identical?"

Department of Statistics, University of Illinois, Urbana-Champaign, April 12, 2012.
- McElroy, Tucker, "On the Computation of Autocovariances for Generalized Gegenbauer Processes."

Department of Statistics Colloquium, The Pennsylvania State University, State College, PA, April 26, 2012.
- Klein, Martin, "Imputation for Nonmonitone Nonresponse in the Survey of Industrial Research and Development."

Seminar, The University of Maryland, College Park, MD, May 3, 2012.

- Slud, Eric, "Parametric Survival Densities from Phase-Type Models," joint with Jiraphan Suntornchost, (Ph.D. student at UMCP)

Annual Meeting of the American Association for Public Opinion Research, Orlando, FL, May 17-20, 2012.

- Bates, Nancy and Mary Mulry, "Did the 2010 Census Social Marketing Campaign Shift Public Mindsets?"

Mini-Workshop on Data Analysis Issues in Statistical Agencies, The University of Hong Kong, Hong Kong, China, June 1, 2012.

- Klein, Martin, "Imputation for Nonmonotone Nonresponse in the Survey of Industrial Research and Development."

2012 International Conference on Establishment Surveys, Montreal, Canada, June 11-14, 2012.

- Mulry, Mary, Broderick Oliver, and Stephen Kaputa, "Study of Treatment of Influential Values in a Monthly Retail Trade Survey."

Federal Committee on Statistical Methodology Data Review Board Information Sharing Group, Washington, DC, June 21, 2012.

- Maria Garcia, "Selective Editing Strategies for the Census Bureau Foreign Trade Statistics Programs."

DC-AAPOR/WSS Summer Conference Preview/Review 2012, Washington DC, June 21-22, 2012.

- Gilary, Aaron, "Small Area Confidence Bounds on Small Cell Proportions in Survey Populations."

2012 Society for Industrial and Applied Mathematics (SIAM) Annual Meeting, Minneapolis, MN, July 9-13, 2012.

- Arceneaux, Taniecea, "Resilience of Small Social Networks with Multiple Relations."

Joint Statistical Meetings, American Statistical Association, San Diego, California, July 28-August 2, 2012.

- Bill Bell, Gauri Datta, and Malay Ghosh, "Benchmarking Small-Area Estimates."
- Chris Blakely, "Coupling X-12 ARIMA-SEATS with a Multidimensional Direct Filter Approach for Signal Extraction in Nonstationary Seasonal Time Series."
- Adrijo Chakraborty and Gauri Datta, "A Hierarchical Mixture Model for Small-Area Estimation."
- Yang Cheng, Patrick Flanagan, and Eric Slud, "Overview of Current Population Survey Methodology."
- Gauri Datta, Abhyuday Mandal, and Anthony Wanjoya, "Small-Area Estimation with Uncertain Random Effects."
- David Findley, "Complementary Properties of an F-Test and an Empirical Spectral Test for Identifying Seasonality in Unadjusted or Seasonally Adjusted Series."
- Scott Holan, and Tucker McElroy, "Flexible Spectral Models for Multivariate Time Series."
- Patrick Joyce, Donald Malec, Roderick Little, and Aaron Gilary, "Application of a Small-Area Model for a Voting Rights Act Tabulation."
- Martin Klein, Peter Linton, and Bimal Sinha, "Tests for Homogeneity of Multinomial Proportions for Sparse Data."
- Jerry Maples, "Estimating the Relative Variance of Replicate Weight Sampling Error Variance Estimators for Rates."
- Thomas Mathew and K. Krishnamoorthy, "Statistical Methodologies for Exposure Data Analysis." Short Course.
- Tucker McElroy, "Subsampling Inference for the Autocovariances of Heavy-Tailed Long-Memory Time Series."
- Brian Monsell, "Evaluating AICC Tests in X-13ARIMA-SEATS."
- Mary Mulry, Broderick E. Oliver, and Stephen Kaputu, "Several Scenarios for Influential Observations in Business Surveys and Methods for Their Treatment."
- Sharon O'Donnell, and Yves Thibaudeau, "Combining Model-Based and Hot-Deck Imputation to Fill Gaps in Longitudinal Surveys."
- Joseph Schafer (delivered by a substitute), Discussant for Survey Sampling and Missing Data Problems – Topic-Contributed Session.

- Eric Slud, Aaron Gilary, and Jerry Maples, “Small-Area Confidence Bounds on Small Cell Proportions in Survey Populations.”
- Andrew Vesper and Ryan Janicki, “Benchmarking Small-Area Estimates: A Minimum Discrimination Information Approach and Other New Perspectives.”
- Jerzy Wieczorek, Ciara Nugent, and Sam Hawala, “A Bayesian Zero-One Inflated Beta Model for Small-Area Shrinkage Estimation.”
- William Winkler, “Estimating Record Linkage Error Rates Without Training Data.”
- Tommy Wright, “The Equivalence of Neyman Optimum Allocation for Sampling and Equal Proportions for Apportioning the U. S. House of Representatives.”

2012 International Total Survey Error Workshop, Santpoort, Netherlands, September 2-4, 2012.
- Mulry, Mary and Bruce Spencer, “A Framework for Empirical Cost Modeling Relating Cost and Data Quality.”

Department of Mathematics and Statistics Colloquium, University of Maryland, Baltimore County; Baltimore, MD, September 7, 2012.
- Klein, Martin, “Statistical Analysis of Noise Multiplied Data Using Multiple Imputation.”

Department of Statistics Colloquium, The University of Georgia, Athens, GA, September 13, 2012.
- Klein, Martin, “Statistical Analysis of Noise Multiplied Data Using Multiple Imputation.”

Department of Mathematics Seminar, The University of Maryland, College Park, MD, September 13, 2012.
- Schafer, Joe, “Flexible Bayesian Models for Process Monitoring of Paradata Survey Quality Indicators.”

Statistics Department Seminar, The George Washington University, Washington, D.C., September 14, 2012.
- Slud, Eric, “Small Area Confidence Bounds on Small Cell Proportions in Survey Populations.”

Department of Mathematics and Statistics Colloquium, University of Maryland, Baltimore County; Baltimore, MD, September 14, 2012.
- McElroy, Tucker, “General and Consistent Signal Extraction for Nonstationary Time Series with Diverse Sampling Rules.”

UN/ECE Work Session on Statistical Data Editing, Oslo, Norway, September 24-26, 2012.
- Garcia, Maria, “An Application of Selective Editing to the US Census Bureau Trade Data.”

Federal Forecasters Conference, Bureau of Labor Statistics, Washington, D.C., September 27, 2012.
- McElroy, Tucker, “Multi-Step Ahead Forecasting of Vector Time Series.”

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

Andy Peytchev, Research Triangle Institute, "Anticipatory Survey Design: Reduction of Nonresponse Bias through Bias Prediction Models," October 25, 2011.

Matthew Turner, The University of Tennessee, Knoxville, "Linear Filtering of Multivariate Time Series Driven by Heavy-Tailed Lévy Noise," November 29, 2011.

Joshua Tokle, University of Washington, "Heat Kernels in Probability and PDE," December 12, 2011.

Rebecca Steorts, (U.S. Census Bureau Dissertation Fellow) University of Florida, "Bayes and Empirical Bayes Benchmarking for Small Area Estimation," December 15, 2011.

Bill Winkler, Bill Yancey, & Ned Porter, CSRM, U.S. Census Bureau, "Record Linkage Course," January 31 & February 1, 2012.

Yves Thibaudeau & Eric Slud, CSRM, U.S. Census Bureau, "Hybrid and Model-Assisted Predictors under Poststratifications of Unknown Size," January 25, 2012.

Chris Blakely, CSRM, U.S. Census Bureau, "iMetric: An Econometric Graphical-User-Interface Software Program for Analyzing Nonstationary Univariate and Multivariate Time Series Data," February 9, 2012.

Darcy Morris, Cornell University, "Joint Modeling of Multivariate Longitudinal Count Data," February 16, 2012.

Parvati Krishnamurty, NORC at the University of Chicago, "Memory Recall of Migration Dates in the NLSY97," February 22, 2012.

Bill Winkler, Bill Yancey, & Ned Porter, CSRM, U.S. Census Bureau, "Record Linkage Course (Second Offering)," March 6-7, 2012.

Yaakov Malinovsky, University of Maryland, Baltimore County, "Prediction of Ordered Random Effects in a Simple Small Area Model," March 8, 2012.

Carolina Franco, University of Maryland, College Park, "Semiparametric Estimation for Kernel Families," March 20, 2012.

Andrew Magyar, Rutgers University, "Design Consistent and Outlier Robust Model Based Estimators," April 17, 2012.

Malay Ghosh, University of Florida, *SUMMER AT CENSUS*, "Design Consistent and Outlier Robust Model Based Estimators," May 15, 2012.

Bill Winkler, CSRM, U.S. Census Bureau, "Edit/Imputation Course," May 17, 2012.

Jay Emerson, Yale University, *SUMMER AT CENSUS*, "Towards High-Performance Computing with R," May 22, 2012.

Ben Klemens, CSRM, U.S. Census Bureau, "A Simulation of Nonresponse and Imputation," May 30, 2012.

Jerry Maples, CSRM, U.S. Census Bureau, "Small Area Variance Modeling of County Poverty Estimates From the American Community Survey," May 31, 2012.

Jian-Guo Liu, Duke University, *SUMMER AT CENSUS*, "Mathematical Analysis of Flocking Dynamics in Biology and Social Science," June 19, 2012.

Alex Stuckey, Australian Bureau of Statistics, "A Comparison of Automated ARIMA Model Selection Methods To Benefit Seasonal Adjustment Using X12-ARIMA," June 21, 2012.

Bill Winkler, CSRM, U.S. Census Bureau, "Edit/Imputation Course (Second Offering)," June 25, 2012.

Bikas Sinha (Retired Professor), Indian Statistical Institute, *SUMMER AT CENSUS*, "Systematic Sampling with Inverse Hypergeometric Stopping Rule For Estimation of Incidence Rate of a Rare Attribute," June 26, 2012.

Ralph Bangs & Larry Davis, University of Pittsburgh, *SUMMER AT CENSUS*, "State of Race in America," June 27, 2012.

Jae-Kwang Kim, Iowa State University, "An Efficient Method of Estimation for Longitudinal Surveys with Monotone Missing Data," June 28, 2012.

Ben Klemens, CSRM, U.S. Census Bureau, "Linux for Data Analysts Class," June 28, 2012.

Sarah Nusser, Iowa State University, *SUMMER AT CENSUS*, "Using Geospatial Information in the Field Survey Operations," July 3, 2012.

Akram Khater, North Carolina State University, *SUMMER AT CENSUS*, "Intersections of Religion and Ethnicity Among Early Middle Eastern Immigrants," July 10, 2012.

Akram Khater, North Carolina State University, *SUMMER AT CENSUS*, "*Film: Cedars in the Pines – The Lebanese in North Carolina*," July 12, 2012.

Sarjinder Singh, Texas A & M University – Kingsville, *SUMMER AT CENSUS*, "A Magical Talk: Estimating at Least Seven Measures of Qualitative Variables from a Single Sample Using Randomized Response Technique," July 17, 2012.

Michael Larsen, The George Washington University, *SUMMER AT CENSUS*, "A Study of Factors Affecting Record Linkage in Federal Statistical Databases," July 18, 2012.

Ann Morning, New York University, *SUMMER AT CENSUS*, "The Nature of Race: How Scientists Think and Teach about Human Difference," July 19, 2012.

Chris Blakely, CSRM, U.S. Census Bureau, "Introduction to *iMetrica* Part II: Bayesian Forecasting, Hybrid Filtering, and Localized Regressors for Time Series," July 19, 2012.

Christopher Bollinger, University of Kentucky, *SUMMER AT CENSUS*, "Two Can Live As Cheaply As One…But Three's A Crowd," July 19, 2012.

Rebecca Andridge, The Ohio State University, *SUMMER AT CENSUS*, "Proxy Pattern-Mixture Analysis for Survey Nonresponse," July 23, 2012.

Matthew Salganik, Princeton University, *SUMMER AT CENSUS*, "Assessing Respondent-Driven Sampling," July 25, 2012.

Matthew Salganik, Princeton University, *SUMMER AT CENSUS*, "Generalized Network Scale-Up Method for Estimating the Sizes of Hard-to-Count Groups: Evidence from Brazil and Rwanda," July 26, 2012.

David Haziza, Université de Montréal, *SUMMER AT CENSUS*, "Robust Estimation in the Presence of Influential Units: A Unified Approach," August 6, 2012.

Peter Brandon, University at Albany, SUNY, *SUMMER AT CENSUS*, "Using the SIPP to Understand the Effectiveness of Welfare Reform," August 7, 2012.

Marc Wildi, Zurich University of Applied Sciences, *SUMMER AT CENSUS*, "Real-Time Forecasting and Signal Extraction: Accuracy, Reliability, and Timeliness," August 8, 2012.

Marlow Lemons, CDAR, U.S. Census Bureau, "n-Cycle Swapping for the American Community Survey," August 9, 2012.

Yazhen Wang, University of Wisconsin-Madison, *SUMMER AT CENSUS*, "Modeling and Analyzing High-Frequency Financial Data," August 14, 2012.

Eric Stone, Temple University, National Agriculture Statistics Service, "Give Your Data a Listen with audiolyzR," August 15, 2012.

Singdhansu Chatterjee, University of Minnesota, *SUMMER AT CENSUS*, "On Using Data-Depth for High-Dimensional Inference in Survey Data Problems," August 15, 2012.

Qian Cai, University of Virginia, *SUMMER AT CENSUS*, "Population Projections: Theory, Practice, and Reflection," August 15, 2012.

Richard Bilsborrow, The University of North Carolina at Chapel Hill, *SUMMER AT CENSUS*, "Collecting Data on International Migrants in the World: Some Issues and Proposals," August 16, 2012

Hishaam Aidi, Columbia University, *SUMMER AT CENSUS*, "The Geography of Race Classification: The Case of Afro-Arab North African Identities," August 21, 2012.

James Berger, Duke University, *SUMMER AT CENSUS*, "Reproducibility of Science: P-values and Multiplicity," August 22, 2012.

Debra Umberson, University of Texas at Austin, *SUMMER AT CENSUS*, "Same Sex Marriage and Health: Fieldwork and Land Mines," August 27, 2012.

Wesley Yung, Statistics Canada, *SUMMER AT CENSUS*, "Some Work on Business Surveys Collection Research at Statistics Canada," August 28, 2012.

Nathan Yau, University of California, Los Angeles, *SUMMER AT CENSUS*, "Visualization that Means Something," August 29 and 30, 2012.

Anthony Damico, Kaiser Family Foundation, "Introduction to R with ACS, CPS, and SIPP Data," September 12, 2012.

Chet Bowie, National Opinion Research Center, *SUMMER AT CENSUS*, "The Science of Management: Implementing Matrix Management in Reimbursable Surveys," September 18, 2012.

Lee De Cola, Data to Insight, "Introduction to R's Spatial Packages," September 25, 2012.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

Arthur S. Fleming Award, Arthur S. Fleming Awards Commission

- **Tucker McElroy** – "…has developed novel statistical methodologies to address significant problems in seasonal adjustment, forecasting, and time series analysis. Seasonal adjustment takes into account recurring seasonal fluctuations, in areas such as retail sales and employment during holiday periods, which are critical to the evaluation of potential government actions to improve the economy. Internationally recognized, Dr. McElroy's published achievements on the theory of signal extraction and model misspecification have been implemented in the X-12-ARIMA software program of the Census Bureau which has been used in central banks and statistical agencies around the world. His recent work has improved X-12-ARIMA's model selection methods and extended it to treat time series data of mingled sampling frequency."

Bronze Medal Award, U.S. Bureau of the Census

- **Chad Eric Russell, … (Team Award) –** "The Data Management System (DMS) provides a secure, controlled environment for sharing and using data within the Census Bureau. Awardees applied their considerable pooled talents to successfully deploying the DMS. The effort was successful due to technical excellence, a comprehensive approach to governance, planning for immediate and continuous training, and creative multi-mode publicity."

- **Patrick Joyce, Don Malec (NCHS), Aaron Gilary, Bill Bell (R&M), … (Team Award) –** "The team developed and implemented small area estimation methods to determine which jurisdictions must provide language assistance to voters per the Voting Rights Act. Using American Community Survey and 2010 Census data, this team demonstrated that estimates that were a weighted combination of direct estimates and "regression type" estimates can be an improvement over the direct estimates alone."

6.2 SIGNIFICANT SERVICE TO PROFESSION

Taniecea Arceneaux
- Refereed a paper for *Statistics and Computing.*

Chandra Erdman
- Career Panelist, Society of Industrial and Applied Mathematics Mid-Atlantic Regional Student Conference, Shippensburg University, Shippensburg, Pennsylvania, April 7, 2012.
- Refereed papers for *Statistics and Computing* and the *Journal of Computational and Graphical Statistics.*

David Findley
- Honored with the Festschrift volume: Bell, W., Holan, S., and McElroy, T. (2012). Economic Time Series: Modeling and Seasonality. New York: Chapman Hall.
- Guest Editor, Special Issue on Seasonal Adjustment, *Taiwan Economic Forecast and Policy*.

Maria Garcia
- Refereed a paper for the *Journal of Official Statistics*
- Session Organizer and Discussant for Topic (vi): "New and Emerging Methods", UN/ECE Work Session on Statistical Data Editing, September 24-26, 2012

Ryan Janicki
- Refereed papers for *Journal of the Royal Statistical Society, Series B* and *Proceedings of Statistics 2011 Canada / IMST 2011 – FIM XX*.

Patrick Joyce
- Refereed papers for *Journal of Survey Methodology* and *Journal of Quantitative Analysis of Sports.*

Martin Klein
- Member, Ph.D. Dissertation in Statistics Committee, University of Maryland, Baltimore County.

Jerry Maples
- Refereed papers for *Computational Statistics and Data Analysis* and *Journal of the American Statistical Association.*

Thomas Mathew
- Associate Editor, *Journal of the American Statistical Association.*
- Associate Editor, *Statistical Methodology.*
- Associate Editor, *Sankhya.*
- Member, American Statistical Association's Committee for the W. J. Youden Award in Interlaboratory Testing.

Tucker McElroy
- Refereed papers for *Computational Statistics and Data Analysis, International Statistical Review, Statistics and Probability Letters, Journal of Time Series Analysis, Applied Stochastic Models in Business and Industry*, and *Journal of the Royal Statistical Society.*
- Organizer, two invited sessions and one topic contributed session, *Joint Statistical Meetings*, 2012.

Mary H. Mulry
- Vice President, American Statistical Association.
- Associate Editor, *Journal of Official Statistics.*
- Member, Program Committee, International Conference on Methods for Surveying and Enumerating Hard-to-Reach Populations, October 31 – November 3, 2012, New Orleans, LA.

Rolando Rodriguez
- Refereed a paper for *Survey Methodology*.

Joe Schafer
- Refereed papers for *Journal of Applied Econometrics, British Journal of Mathematics and Statistics,* and *Psychological Methods.*

Eric Slud
- Refereed papers for *Journal of Statistical Planning and Inference, Statistical Methodology, The Open Journal of Probability and Statistics*, *National Science Foundation,* and *Journal of Official Statistics.*
- Associate Editor, *Lifetime Data Analysis.*
- Associate Editor, *Journal of the Royal Statistical Society, Series B (Methodological).*
- Member, *Steering Committee for National Research Council* Workshop on "Future Directions for the NSF National Patterns of Research and Development" funded by the National Center for Science and Engineering Statistics (NCSES).
- Reviewer, External Progress Report on "Pilot Studies Testing Off-Site Survey Designs for Estimating Total Marine Recreational Fishing Effort" (May 2012, with Patrick Joyce) for the Office of Science and Technology NOAA Fisheries Service (National Marine Fisheries Service).
- Member, National Academy of Sciences Steering Committee of Future Directions for the National Science Foundation National Patterns of Research and Development.
- Reviewer, NRC Final Report of a CNSTAT Panel on Estimating Children Eligible for School Nutrition Programs Using the ACS.

William Winkler
- Refereed papers for *Information Systems* and papers for *Statistical Data Protection 2012*.

- Reviewed a National Science Foundation (NSF) proposal.
- Associate Editor, *Journal of Privacy and Confidentiality.*
- Associate Editor, *Transactions on Data Privacy.*
- Member, Program Committee, Statistical Data Protection 2012.

Tommy Wright
- Associate Editor, *The American Statistician*.
- Member, Fellows Committee, American Statistical Association.
- Member, Advisory Board, Department of Mathematics and Statistics, Georgetown University.
- Member, Workgroup on Master's Degrees, American Statistical Association.

Derek Young
- Refereed papers for *Entropy*, *Journal of Nonparametric Statistics, Journal of Applied Statistics, Neural Computation*, and *Wiley Interdisciplinary Reviews: Computational Statistics.*

6.3 PERSONNEL NOTES

Rob Creecy retired from the Census Bureau after 39 years of Federal Service.

Elizabeth Huang retired from the Census Bureau after 35 years of Federal Service.

Gloria Prout retired from the Census Bureau after 38 years of Federal Service.

Julie Tsay retired from the Census Bureau after 22 years of Federal Service.

Lynn Weidman retired from the Census Bureau after 32 years of Federal Service.

Sarah Wilson received her M.S. degree in linguistics from Georgetown University and accepted a position as Editor (Scientific) with our center.

Osbert Pang joined our Time Series Research Group.

Taniecea Arceneaux joined our Statistical Computing Applications and Data Visualization Research Group as a Postdoctoral Researcher.

Douglas Galagate (graduate student in mathematical statistics at University of Maryland, College Park) joined our center as an intern.

Joshua Tokle joined our Machine Learning & Computational Statistics Research Group.

Jerzy Wieczorek (who has been a member of the Social, Economic & Housing Statistics Division for slightly more than two years) was reassigned to our center and joined our Small Area Estimation Research Group.

Erica Magruder (a new member of the Human Resources Division's Mixed-Tour Program) joined our Mathematical Statistics Area as a Research Assistant.

Andrew Magyar joined our Sampling and Estimation Research Group as a Postdoctoral Researcher.

Gauri Datta (member of the statistics faculty at the University of Georgia) began (late FY 2011) a Schedule A Appointment in our Small Area Estimation Research Group.

Ben Klemens became the leader of our Statistical Computation and Data Visualization Research Group.

Jerry Maples became the leader of our Small Area Estimation Research Group.

Summer Visitors:

- Joel Beard joined our Experimentation, Simulation, and Modeling Research Group as an intern.
- Derrick Simmons (junior majoring in mathematics at Howard University) joined our Experimentation, Simulation, and Modeling Research Group as an intern.
- Laura Becht (sophomore at the College of William and Mary) joined our center as an intern.
- Charles (Eddie) Davis (junior majoring in mathematics and computer science at the University of Maryland, College Park) joined our Experimentation, Simulation, & Modeling Research Group as an intern.
- Barrett Veazey (junior majoring in mathematics at Georgetown University) joined our Experimentation, Simulation, & Modeling Research Group as an intern.
- John Zylstra (Ph.D. student in statistics at the University of Maryland-Baltimore County) joined our Missing Data Methods Research Group as an intern.

Alisha Armas joined our Statistical Computing Area as a research assistant.

Carolina Franco joined our Sampling & Estimation Research Group.

Darcy Morris joined our Missing Data Methods Research Group.

APPENDIX A **Center for Statistical Research and Methodology FY 2012**
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		
5610202 / 6510206	Statistical Design and Estimation		
	1. *Decennial Record Linkage*	William Winkler	Tom Mule
	2. *Voting Rights Section 203 Model Based Methodology: Research, Development, and Production*	Patrick Joyce	Mark Asiala
	3. *Synthetic Decennial Microdata File*	Martin Klein	Laura Zayatz
	4. *Coverage Measurement Research*	Jerry Maples	Pat Cantwell
	5. *Accuracy of Coverage Measurement*	Mary Mulry	Pat Cantwell
6510201	Coding, Editing, and Imputation Study		
	6. *Software Development (TEA)*	Ben Klemens	Pat Cantwell
	7. *Software Analysis and Evaluation*	Rolando Rodriguez	Andrew Keller
6710201	Enhancing Demographic Analysis		
	8. *Enhancing Demographic Analysis for the 2020 Census*	Ben Klemens	Jason Devine
	9. *Privacy and Confidentiality for the 2020 Census*	Martin Klein	Jennifer Childs
6810205	Matching Process Improvement		
	10. *2020 Unduplication Research*	Michael Ikeda	Craig Johnson
TBA	Statistical Design for 2020 Planning, Experimentation, and Evaluations		
	11. *Master Address File (MAF) Error Model and Quality Assessment*	Derek Young	Robin Pennington
	12. *Supplementing and Supporting Non-Response with Administrative Records*	Michael Ikeda	Randall Neugebauer
5385260	American Community Survey (ACS)		
	13. *ACS Applications for Time Series Methods*	Tucker McElroy	Mark Asiala
	14. *ACS Imputation Research and Development*	Yves Thibaudeau	Deborah Griffin
	DEMOGRAPHIC		
TBA	Demographic Statistical Methods Division Special Projects		
	15. *Tobacco Use Supplement (NCI) Small Domain Models*	Aaron Gilary	Benmei Liu
0906/1442	Demographic Surveys Division Special Projects		
	16. *Data Integration*	Ned Porter	Marie Pees
7323008/ 7523012	National Crime Victimization Survey		
	17. *Analyzing the Effects of Sample Reinstatement, Refresher Training Experiment, and Process Monitoring and Fitness for Use*	Joe Schafer	Bill Samples
TBA	Population Division Projects		
	18. *Population Projections*	Tucker McElroy	Jennifer Ortman
1465444	Survey of Income and Program Participation (RE-SIPP) Research		
	19. *Model-Based Imputation for the Demographic Directorate*	María García	Martha Stinson
7165000	Social, Economic, and Housing Statistics Division (SEHSD) Small Area Estimation Projects		
	20. *Research for Small Area Income and Poverty Estimates (SAIPE)*	Jerry Maples	Wes Basel
	21. *Small Area Health Insurance Estimates (SAHIE)*	Ryan Janicki	Don Luery
1442555	Improving Poverty Measures/IOE		
	22. *Tract Level Estimates of Poverty from Multi-year ACS Data*	Jerry Maples	Wes Basel
	ECONOMIC		
2320254	23. *Investigation of Selective Editing Procedures for Foreign Trade Programs*	María García	Ryan Fescina
2320252	Time Series Research		
	24. *Seasonal Adjustment Support*	Brian Monsell	Kathleen McDonald Johnson
	25. *Seasonal Adjustment Software Development and Evaluation*	Brian Monsell	Kathleen McDonald Johnson
	26. *Research on Seasonal Time Series - Modeling and Adjustment Issues*	Tucker McElroy	Kathleen McDonald Johnson
	27. *Supporting Documentation and Software for X-12-ARIMA and X-13A-S*	Brian Monsell	Kathleen McDonald Johnson
TBA	28. *Governments Division Project on Decision-Based Estimation*	Eric Slud	Bac Tran

APPENDIX B

FY 2012 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 eleventh year to provide *seven measures of performance*, this time for the fiscal year 2012. For FY 2012, the *measures of performance* for our center are:

- *Measure 1. Overall, Work Met Expectations:* Percent of FY 2012 Program Sponsored Projects/Subprojects where sponsors reported that work met their expectations.
- *Measure 2. Established Major Deadlines Met:* Percent of FY 2012 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 2012 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight.
- *Measure 3b. Plans for Implementation*: Of the FY 2012 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 2012 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 2012.
- *Measure 6. Proceedings Publications:* Number of proceedings publications documenting research that appeared in FY 2012.

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 November 26 thru December 7 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 2012 ***(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 2012 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 2012 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 2012 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

(CSRM Chief will coordinate last two signatures as noted.)

Third_____
Sponsor Division Chief Signature Date

Fourth_____
CSRM Center Chief Signature Date

Center for Statistical Research and Methodology

Research & Methodology Directorate

STATISTICAL COMPUTING AREA

Joe Schafer
- Alisha Armas

Machine Learning & Computational Statistics Research

Bill Winkler
- Joshua Tokle
- William Yancey

Statistical Computing Applications & Data Visualization Research

Ben Klemens
- Taniecea Arceneaux (Postdoc)
- Tom Petkunas
- Ned Porter
- Rolando Rodriguez

Missing Data Methods Research

Yves Thibaudeau
- Chandra Erdman
- Douglas Galagate (S)
- Maria Garcia
- Martin Klein
- Darcy Morris
- Jun Shao (U. of WI)
- John Zylstra (S)

Research Computing Systems

Chad Russell
- VACANT

MATHEMATICAL STATISTICS AREA

Eric Slud
- Erica Magruder (HRD)

Sampling & Estimation Research

Eric Slud (Acting)
- Carolina Franco
- Mike Ikeda
- Patrick Joyce
- Mary Mulry
- VACANT

Small Area Estimation Research

Jerry Maples
- Gauri Datta (U. of GA)
- Aaron Gilary
- Ryan Janicki
- Jerzy Wieczorek
- VACANT

Time Series Research

Brian Monsell
- Chris Blakely (Postdoc)
- David Findley
- VACANT

Tucker McElroy
- Osbert Pang

Experimentation, Simulation, & Modeling Research

Tommy Wright (Acting)
- Thomas Mathew (UMBC)
- Derrick Simmons (S)
- Derek Young

Tommy Wright, Chief
- Kelly Taylor
- Ann Dimler
- Laura Becht (S)
- Margo Anderson (F)
- Bill O'Hare (F)
- Michael Leibert
- Michael Hawkins
- Esan Summer (S)
- Sarah Wilson

(F) ASA/NSF/Census Research Fellow
(S) Student

September 30, 2012