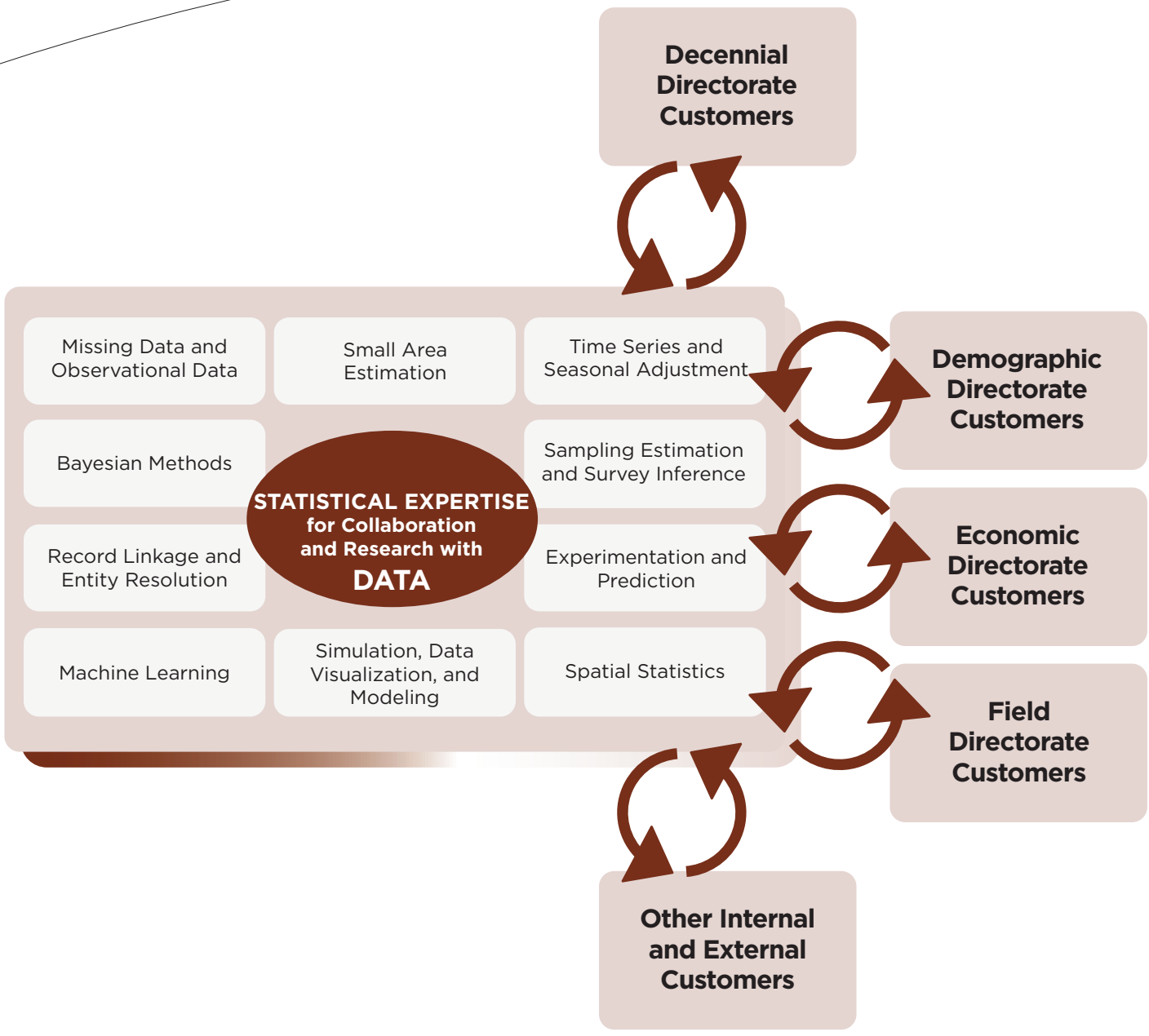
Annual Report of the Center for Statistical Research and Methodology

Research and Methodology Directorate

Fiscal Year 2023



United States® Census Bureau

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Since August 1, 1933—

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

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

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

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

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

---

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

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

We help the Census Bureau improve its processes and products. For fiscal year 2023, this report is an accounting of our work and our results.

Center for Statistical Research & Methodology
https://www.census.gov/topics/research/stat-research.html

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 Census Bureau 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 2023 follow, and more details are provided within subsequent pages of this report:

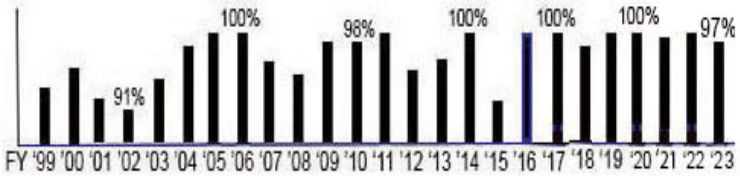
- CSRM researchers worked with Decennial Directorate to develop a prototype of proposed methodology for statistical model-guided Mobile Questionnaire Assistance (MQA) placement; a utility function was developed to express the predicted effect of adding or removing a placement - at a particular location and time - using modeled association between MQA exposure and response rate. [CSRM (Raim); DCMD (Moore); DSSD (Fletcher)]
- In support of the Continuous Count Study which is aimed at the production of an administrative records enumeration of the national population, CSRM researchers examined and ran some basic cross-tabulations on the Demographic Frame, made comparisons with the 2020 Census Edited File (CEF), and produced a draft report documenting some of the exploratory results. [CSRM (Ikeda); R&M (Mule)]
- CSRM and Social, Economic, & Housing Statistics Division researchers continued work on a research report on a Geographically Weighted Regression (GWR) Small Area Model which brings the ideas of GWR into the Fay-Herriot small area model framework; researchers developed a graphical method to detect when parameters vary across small areas compared to a singular model where parameters are the same for all areas. [CSRM (Maples, Dompreh); SEHSD (Basel)]
- CSRM and Economic Directorate researchers worked to complete a research paper on a hierarchical Bayesian mass imputation methodology for estimating state-level retail sales based on data from a third-party aggregator. [CSRM (Morris); ESMD (Kaputa, Thompson); EID (Hutchinson)]
- CSRM and Economic Statistical Methods Division researchers developed and implemented custom code to investigate the performance of multivariate seasonal adjustments compared to univariate adjustments; initial findings suggest that when correctly specified, the multivariate approach outperforms the univariate approach. [CSRM (Livsey, Pang); ESMD (Viehdorfer)]
- Joining an Economic Directorate effort in support of the application of small area estimation methods to the Annual Integrated Economic Survey (AIES), CSRM and Economic Directorate researchers explored preliminary linear Fay-Herriot models with a chief goal of ultimately producing state-level estimates for all AIES core items by three-digit North American Industry Classification System (NAICS3) groups. [CSRM (Maples, Datta, Aleshin-Guendel, Janicki); ESMD (Thompson, Kaputa); EMD (Maison)]
- With colleagues in the Center for Optimization & Data Science and the Computer Services Division, CSRM staff continued to maintain and support the Integrated Research Environment (IRE) and the Cloud Research Environment (CRE) prototype, and planned and tested a migration strategy to take the existing IRE cluster and move it to a new cluster ("IRE New") with a new OS (RHEL 8), a new graphical user interface (XFCE), a new version of PBS, and newer versions of many statistical packages (Matlab, Mathematica, etc.). [CSRM (Russell); CODS (Damineni)]
- CSRM researchers reported simulation results using the E-M Algorithm to fit the matching parameters which are used to calculate the match weights which are values between 0 and 1; the E-M Algorithm might realize convergence to the extreme values of 0 or 1, hence possibly creating over confidence in assigning a match status; one result from the simulation was that when the design matrix is reduced from 3 parameters to 2 parameters, convergence to 0 or 1 was completely avoided while estimating the true parameters reasonably well. [CSRM (Weinberg, Thibaudeau)]
- CSRM researchers updated the two visuals of The Ranking Project based on ACS data which include "Comparisons of A State with Each Other State" and "Estimated Rankings of All States" with 1-Year ACS data on 89 Topics for 2021; the two updated visualizations now provide ACS data for 2018, 2019, and 2021 and can be accessed by googling "The Ranking Project." [CSRM (Wright, Hall, Yau); Colby College (Wieczorek)]
- CSRM and Research & Methodology researchers developed new methodology for automatic clustering of data for the purpose of producing accurate estimates with an emphasis on model-based methods using data collected under an informative sample survey design; the distributional theory was fully derived and code was developed. [CSRM (Janicki, Parker); R&M (Holan)]
- Research & Methodology and CSRM researchers made progress on several time series projects; (a) writing of text and code for a book on multivariate real-time seasonal adjustment and forecasting, (b) explored point process model of seasonal data; and (c) refined new seasonality diagnostic based upon forecast errors and forecast revisions. [R&M (McElroy); CSRM (Livsey, Pang, Roy)]
- CSRM researcher published a manuscript on direct sampling with step function which features several applications; the direct sampler (introduced by Walker et al, 2011) may be used to draw the unobserved statistics as latent random variables within a Gibbs sampler. [CSRM (Raim)]

How Did We[1] Do...

For the 25th year, we received feedback from our sponsors. Near the end of fiscal year 2023, our efforts on 33 of our programs (Decennial, Demographic, Economic, External, etc.) 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 33 questionnaires were obtained with the following results (The graph associated with each measure shows the performance measure over the last 25 fiscal years):

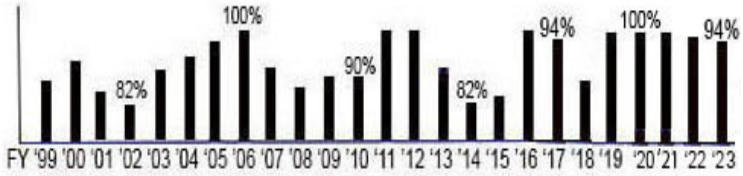
Measure 1. Overall, Work Met Expectations

Percent of FY2023 Program Sponsored Projects/Subprojects where sponsors reported that overall work met their expectations (agree or strongly agree) (32 out of 33 responses) ….. 97%



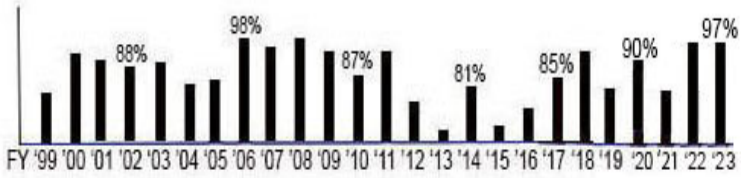
Measure 2. Established Major Deadlines Met

Percent of FY2023 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met (16 out of 17 responses) ……..…..…..….. 94%



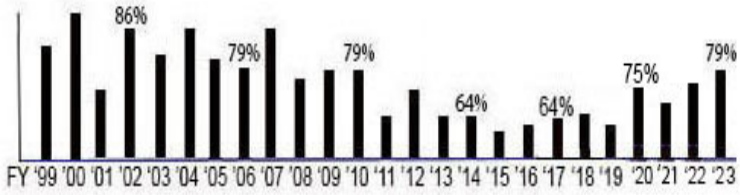
Measure 3a. At Least One Improved Method, Developed Technique, Solution, or New Insight

Percent of FY2023 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight (32 out of 33 responses) … 97%



Measure 3b. Plans for Implementation

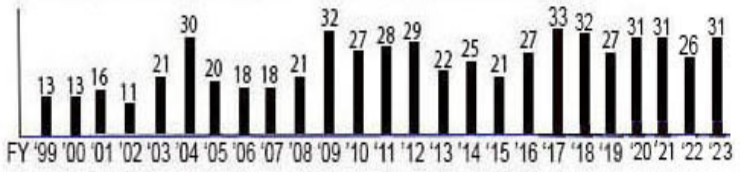
Of these FY2023 Program Sponsored Projects/Subprojects reporting at least one improved method, technique developed, solution, or new insight, the percent with plans for implementation (26 out of 33 responses) …….. 79%



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

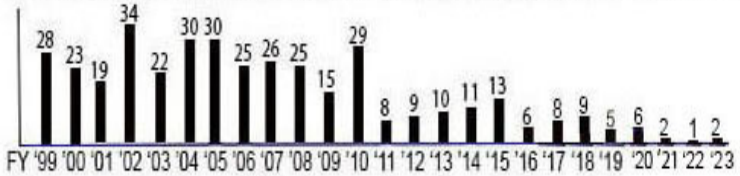
Measure 4. Journal Articles (Peer-Reviewed), Publications

Number of peer reviewed journal publications documenting research that appeared (23) or were accepted (8) in FY2023 ……... 31



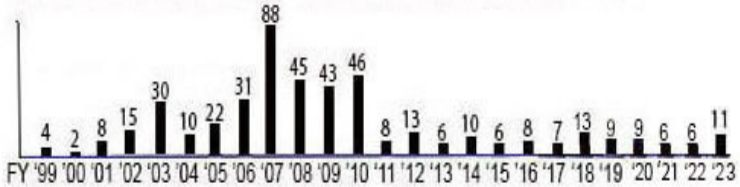
Measure 5. Proceedings, Publications

Number of proceedings publications documenting research that appeared in FY2023 …….. 2



Measure 6. Center Research Reports/Studies, Publications

Number of center research reports/studies publications documenting research that appeared in FY2023 … 11



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**
Decennial Directorate 1
1.1 Project 6550J06 – Redistricting Data Program
1.2 Project 6550J08 – Data Products Dissemination Prep/Review/Approval
1.3 Project 6650J01 – PES Planning & Project Management
1.4 Project 6650J20 – 2020 Evaluations – Planning & Project Management
1.5 Project 5350J01 – Address Frame Updating Activities
1.6 Project 5350J04 – Demographic Frame Updating Activities
1.7 Project 5450J06 – Content, Forms Design, and Language Project
1.8 Project 5450J10 – In-Person Enumeration Planning and Support
1.9 Project 5450J20 – In-Office Enumeration Planning and Support
1.10 Project 5450J21 – Response Data Quality
1.11 Project 5450J23 – Response Processing Planning and Support
1.12 Project 5550J01 – Data Products Creation & Dissemination
1.13 Project 5650J02 – PES Planning and Project Management
1.14 Project 6385J70 – American Community Survey

Demographic Directorate 10
1.15 Project TBA – Demographic Statistical Methods Division Special Projects
1.16 Project 0906/1444X00 – Demographic Surveys Division (DSD) Special Projects
1.17 Project 7165023 – Social, Economic, & Housing Statistics Division Small Area Estimation Projects

Economic Directorate 11
1.18 Project 1183X01 – General Economic Statistical Support
1.19 Project 1183X90 – General Economic Statistical Program Management

Census Bureau 14
1.20 Project 0331000 – Program Division Overhead
1.21 Project 9401021 – Modeling National Cancer Center Tobacco Use Supplement/Current Population Survey Outcomes

2. RESEARCH **16**
2.1 Project 0331000 – General Research and Support
Missing Data & Observational Data Modeling
Record Linkage & Machine Learning
Sampling Estimation & Survey Inference
Small Area Estimation
Spatial Analysis and Modeling
Time Series & Seasonal Adjustment
Experimentation, Prediction, & Modeling
Simulation, Data Science, & Visualization
SUMMER AT CENSUS
Research Support and Assistance

3. PUBLICATIONS **32**
3.1 Journal Articles (Peer-Reviewed), Publications
3.2 Books/Book Chapters
3.3 Proceedings Papers
3.4 Center for Statistical Research & Methodology Research Reports
3.5 Center for Statistical Research & Methodology Study Series
3.6 Other Reports

4. TALKS AND PRESENTATIONS **35**

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES **37**

6. PERSONNEL ITEMS **39**
6.1 Honors/Awards/Special Recognition
6.2 Significant Service to Profession
6.3 Personnel Notes

APPENDIX A
APPENDIX B

1. COLLABORATION

1.1 REDISTRICTING DATA PROGRAM (Decennial Project 6550J06)

1.2 DATA PRODUCTS DISSEMINATION PREPARATION/REVIEW/APPROVAL (Decennial Project 6550J08)

1.3 PES PLANNING & PROJECT MANAGEMENT (Decennial Project 6650J01)

1.4 2020 EVALUATIONS – PLANNING & PROJECT MANAGEMENT (Decennial Project 6650J20)

1.5 ADDRESS FRAME UPDATING ACTIVITIES (Decennial Project 5350J01)

1.6 DEMOGRAPHIC FRAME UPDATING ACTIVITIES (Decennial Project 5350J04)

1.7 CONTENT, FORMS DESIGN, & LANGUAGE (Decennial Project 5450J06)

1.8 IN-PERSON ENUMERATION PLANNING & SUPPORT (Decennial Project 5450J10)

1.9 IN-OFFICE ENUMERATION PLANNING & SUPPORT (Decennial Project 5450J20)

1.10 RESPONSE DATA QUALITY (Decennial Project 5450J21)

1.11 RESPONSE PROCESSING PLANNING & SUPPORT (Decennial Project 5450J23)

1.12 DATA PRODUCTS CREATION & DISSEMINATION (Decennial Project 5550J01)

1.13 PES PLANNING & PROJECT MANAGEMENT (Decennial Project 5650J02)

A. Summary of Administrative Records Modeling for Some Enumerations in the 2020 Census

Description: The 2020 U.S. Census is the first U.S. Census to use administrative records (ARs) to enumerate some households. Previously, staff collaborated to prepare a high-level discussion of the research and methodology underlying the use of ARs in the enumeration of households living in housing units at some addresses in Nonresponse Follow-up (NRFU) while maintaining the quality of the data and reducing the cost of NRFU. The topics include: (1) a brief introduction to administrative records, (2) a description of the research and development that occurred from 2012 through 2018 to prepare for using administrative records in census enumeration, (3) the original plan for using ARs in enumeration, (4) the modifications and adaptations required to cope with the unforeseen disruptions in the implementation of 2020 U.S. Census due to the pandemic. Throughout the document, the descriptions of the research and methodology include the rationale behind the resulting decisions.

Highlights: During FY 2023, staff collaborated with staff in the Decennial Statistical Studies Division (DSSD) to prepare the paper to submit to a journal. The paper shows how the Census Bureau incorporated the use of administrative records and other innovations into the 2020 Census. One major innovative use of administrative records was to enable classifying some addresses as occupied, vacant, or nonresidential when neither a self-response nor non-response follow-up response was available. Currently, staff members are awaiting the journal's decision.

Staff: Mary Mulry (682-305-8809)

A.1 Comparisons of Administrative Records Rosters to Census Self-Responses and NRFU Household Member Responses

Description: The Bureau of the Census Scientific Advisory Committee recommended that the Census Bureau conduct analyses that compared census rosters and administrative records (AR) rosters for addresses where both types of rosters were available. As suggested, the Census Bureau has initiated a study that focuses on addresses where both a census roster and an AR roster are available, but the two rosters differ on the size of the household. The study is restricted to addresses where the census roster is a self-response or a Nonresponse Follow-up (NRFU) household member response since these are the highest quality responses. Of particular interest is the situation where the census roster lists one more or one less person than the administrative records identify as residing at the address. When an address had both an AR

roster and a census self-response or a NRFU household member response, the response submitted by the household was the one that was used for the census enumeration in most circumstances.

Highlights: During FY 2023, staff in our Center for Statistical Research & Methodology (CSRM) collaborated with staff in the Decennial Statistical Studies Division (DSSD) and Center for Economic Studies (CES) to prepare a paper comparing 2020 Census rosters from self-responses and non-response follow-up household member responses to administrative rosters at addresses where both are available, and they differ on the household size. One finding from the study indicates that both the mode of response and the amount of recall required of the respondent due to the length of time since April 1 affect the agreement rate between the census roster and the administrative records roster. The team submitted the paper to a journal during FY2023 and is waiting to receive the journal's decision.

Staff: Mary Mulry (682-305-8809)

A.2 Advances in the Use of Capture-Recapture Methodology in the Estimation of U.S. Census Coverage Error

Description: The methodology used to evaluate the coverage of the decennial census has evolved over the years. The methodology and estimation of net coverage error in the 2010 Census that produced the estimates of census coverage error relied on the 2010 Post Enumeration Survey (PES). The data collection methods in 2010 included new quality control procedures and an estimation approach that differed from the estimation method used in the prior PES programs conducted from 1980 through 2000. The implementation of the 2020 PES used essentially the same methodology for data collection and estimation as that employed for the 2010 PES. However, the COVID-19 pandemic resulted in some unexpected delays in the 2020 PES data collection and processing. Staff will produce a paper to document some of this work.

Highlights: During the first part of FY 2023, the book, *Recent Advances on Sampling Methods and Educational Statistics* was published by Springer. The book contained an invited paper co-authored by staff in CSRM and DSSD with title "Advances in the Use of Capture-Recapture Methodology in the Estimation of U.S. Census Coverage Error," which has been well received. The book is in honor of S. Lynne Stokes on her retirement from Southern Methodist University in the summer of 2022. Lynne is a former employee of the Census Bureau who contributed to the methodology for data collection and estimation for the post Enumeration Survey (PES) at different points in her career. One of her contributions involved the development of an estimator of the residential fabrication in the final data from the PES interviews. This project is complete.

Staff: Mary Mulry (682-305-8809)

B. Supplementing and Supporting Nonresponse with Administrative Records

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

Highlights: During FY 2023, staff wrote three draft memoranda and made minor revisions to two additional draft memoranda documenting results of multinomial logistic regression models fit on 2010 administrative records (AR) housing unit (HU) data for non-response follow-up (NRFU) units (with 2010 Census Unedited File (CUF) HU size as the dependent variable) and applied to the 2020 AR modeling HU data. The three new draft memoranda are for: a model based on AR household size (with separate submodels for different values), a "combined" model using IRS 1040 household count but without separate submodels for different values, and a model based on IRS 1040 household count and then AR household size. Staff also revised the draft memoranda for the "initial" model (based on IRS 1040 household count) and the model based on AR household composition. Staff wrote three draft memoranda comparing outlier detection results to the 2020 AR modeling results and the 2020 CUF. The first looked at outlier detection results for AR Modeling "Status Not Assigned" units not assigned by AR by CUF Final Status. The second did a similar comparison for AR Modeling Deletes not assigned by AR by CUF Final Status. The third did a similar comparison for AR Modeling Closeout Deletes not assigned by AR by CUF Final Status. For AR Modeling "Status Not Assigned" not assigned by AR, units with CUF Final Status of Delete tend to have somewhat higher ratio object scores and distributions of individual modeling variables shifted toward values suggesting questionable unit status. For AR Modeling Deletes not assigned by AR and AR Modeling Closeout Deletes not assigned by AR, the distributions of both ratio object score and individual modeling variables tend to be reasonably similar regardless of CUF Final Status.

Staff: Michael Ikeda (x31756)

C. 2020 Census Privacy Variance

Description: The Census Bureau is investigating the within run variance of the 2020 Census differential privacy (DP) algorithm. Specifically trying to identify the accuracy by which individual counts can be estimated given differing levels of released data. For a fixed privacy budget, this project treats true counts as unknowns and estimates them from the released

differentially private data. This surmounts to understanding and solving what is known as a least absolute deviations regression. The ultimate objective is to explore via simulation the possibility of economizing on computation, to approximate the desired variance by actually computing simulated variances in slightly simpler problems with fewer marginal-total related observations.

Highlights: During FY 2023, staff made significant improvements in computation time achieved by decoupling L1 regression runs. This innovative approach involved treating individual margins and cells separately, effectively reducing the dimensionality of the problem and enabling parallel processing of subsets for increased computational efficiency. Subsequently, efforts were directed toward code refactoring to enhance clarity and flexibility. The code is now more readable and logically structured, with increased modularity facilitating independent runs of different code segments. Additional functionality was incorporated to digest run outputs, aiding in progress tracking and issue identification. Overall, these advancements underscore the Census Bureau's commitment to refining and optimizing the differential privacy algorithm for the 2020 Census.

Staff: James Livsey (x33517), Eric Slud

D. Experiment for Effectiveness of Bilingual Training
Description: Training materials were available for enumerators in the 2020 Census to communicate with non-English speaking households. Previously, such situations were left to the enumerator's discretion, and intended census messaging may not have been conveyed uniformly. The Census Bureau would like to measure the effect of this new training on response rate and other key metrics. The goal of this project is to prepare and analyze results from a statistical experiment embedded in the census, subject to operational constraints such as dynamic reassignment of cases and the potential for both trained and untrained enumerators to visit the same households.

Highlights: During FY 2023, staff gathered data from the Enterprise Data Lake, established a coding for outcomes, and formed criteria for inclusion into the experimental analysis. Continuation-ratio logit models were fit to nonresponse follow-up (NRFU) data to associate response rates with factors in the experiment, including whether enumerators are bilingual, whether they receive the new training, stage of NRFU (contact attempt 1, 2, or 3), and the area census office (ACO) associated with the attempt. A simultaneous test based on Holm's method was proposed for the union-intersection hypothesis that the new training does not lower response rates on any of the first three attempts. Estimates of odds-ratios and associated confidence intervals were prepared to compare the odds of a response for enumerators given the new training versus those not given the new training. Preparation of results into a report is in progress.

Staff: Andrew Raim (x37894), Kimberly Sellers, Renee Ellis (CBSM), Mikelyn Meyers (CBSM)

E. Unit-Level Modeling of Master Address File Adds and Deletes
Description: This line of research serves as part of the 2020 Census Evaluation Project on Reengineered Address Canvassing authored by Nancy Johnson. Its aim is to mine historical Master Address File (MAF) data with the overall goal of developing a unit-level predictive model by which existing MAF units may be added or deleted from the current status of live residential housing units for purpose of sampling (e.g., in the American Community Survey sampling universe) or decennial census coverage. There has never been such a predictive model at unit level, nor a concerted effort to mine historical unit-level MAF records for predictive information, and the search for such a unit-level model promises new insights for which MAF units outside the filtered HU universe are most likely to (re-)enter that universe, and also might suggest useful ways to decompose the MAF population in assessing the effectiveness of in-office canvassing procedures.

Highlights: During FY 2023, staff collected and extracted data resources on the Integrated Research Environment (IRE) including MAFX, ACS housing data, and CoreLogic Tax and DEEDS data relevant to addition and deletion of MAFIDs from the ACS sampling housing-unit universe. The goal was first to re-do and then extend predictive analyses using data from 2015 and earlier to predict whether existing 2015 units would change their ACS Universe status in 2016. These status changes were rare events and difficult to predict, so the techniques used were primarily based in decision trees (recursive partitioning and Random Forests). These analyses, performed and validated on several states' data, will result in a final report in Fall 2023 as part of the Decennial 2020 evaluation project on Targeted Address Canvassing.

Staff: Eric Slud (x34991), Nancy Johnson (DSSD)

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

Highlights: During FY 2023, staff assisted the Coverage Estimation Using Administrative Records subteam of the Continuous Count Study program, led by Tom Mule. Staff have been reviewing dual system estimators, using a log-linear model framework, for population estimates.

Staff has also provided feedback on methodologies and exploratory analyses of population counts using

administrative records, specifically the Demographic Frame dataset. The team is focusing on the subset of people that have multiple records at different physical addresses. Staff is researching on where they should be counted, or alternatives for handling those cases in model fitting and prediction.

Staff: Jerry Maples (x32873), Ryan Janicki

G. Empirical Investigation of the Minimum Total Population of a Geographic District to Have Reliable Characteristics of Various Demographic Groups

Description: A key message from earlier empirical work on variability of the TopDown Algorithm (TDA) is that variability in the TDA increases as we consider decreasing levels of geography and population (especially for certain subpopulations). That is, it is the smaller geographic districts with smaller populations where we observed more variability when comparing swapping (SWA) results with TDA results using 2010 Census data. This project is an attempt to take a closer look, using statistical modeling, at variability for smaller districts and to seek an answer to the following question: "What is the minimum Total (ideal) population of a district to have reliable characteristics of various demographic groups?"

Highlights: During FY 2023, staff investigated the use of statistical models to predict reliability of block group population counts produced by the TopDown Algorithm. Staff identified appropriate covariate data sources, including the American Community Survey and prior decennial census counts. Staff developed appropriate metrics for empirically evaluating the statistical models and showed that the Bayesian Additive Regression Trees (BART) produces estimates that are more accurate as compared to linear regression. Staff created a manuscript draft which details the findings for 2010 PL94-171 data.

Staff: Kyle Irimata (x36465), Tommy Wright

H. Comparing Swapping Records Results with Differential Privacy Records Results

Description: Under this project, staff will compare results from two algorithms that added noise to the Census Edited File of the 2010 Census. The first algorithm applied noise using a Swapping technique (SWA) and released data to the public known as Summary File 1 (SF1). The second algorithm applied noise using a Differentially Private technique and is part of the 2020 Census Disclosure Avoidance System (DAS).

Highlights: During FY 2023, a procedure was developed to create tables from the 2010 Census Edited File that have been treated with a Swapping Algorithm (SWA/SF1) and tables from the 2010 DHC Data Demonstration system. These tables were then compared with results developed by the data.census.gov advanced search system. This procedure was broadened to allow the user to create Oracle databases to query the data to create comparative tables for every block in the country.

The SWA/SF1 tables and results were also compared with results developed by the tidycensus system. This project is complete.

Staff: Tom Petkunas (x33216), Joseph Engmark, Tommy Wright

I. Statistical Modeling to Augment 2020 Disclosure Avoidance System

Description: Public data released from the decennial census consists of a number of contingency tables by race, geography, and other factors such as age and sex. These data can be found in products such as the Public Law 94-171 (PL94) Summary File, Summary Files 1 and 2, and the American Indian and Alaska Native (AIAN) Summary File, at varying levels of granularity. Tables involving detailed race and/or geography, crossed with other factors, can pose a risk of unintended disclosure of confidential respondent data. A disclosure avoidance system (DAS) utilizing differential privacy (DP) is being prepared for the release of 2020 decennial census data. However, more detailed data generally require more noise to ensure that a given level of privacy is satisfied; this in turn decreases the utility of the data for users. This project explores a hybrid approach where core tables are released via the DAS and statistical models are used to produce more granular tables from DAS output. In particular, models which capture characteristics jointly across tables and account for spatial structure will be considered.

Highlights: During FY 2023, staff engaged with stakeholders in a series of experiments to evaluate model-based tabulations. Estimates and associated intervals were compared to privacy-protected noisy measurements and underlying Census edited File (CEF) counts. Staff proposed models for state and national tabulations in the Supplementary Demographic Housing and Characteristics (S-DHC) File based on the truncated multivariate normal distribution. Tabulations prepared under these models were demonstrated to enforce basic constraints, such as counts being non-negative and average household sizes being no less than 1, which may be important to eventual users of the data product. For tabulations based on smaller geographies (such as counties) staff proposed models utilizing linear regression with lognormal outcomes and additive noise due to the privacy protection mechanism. To construct design matrices, a mapping between factors in outcomes and auxiliary data based on past census and ACS tabulations was considered. Staff collaborated with MITRE consultants to develop enterprise-level software - which can be maintained and operated by the enterprise - from

the initial research code. Staff published a manuscript on spatial change-of-support modeling of noisy measurements.

Staff: Andrew Raim (x37894), Ryan Janicki, Kyle Irimata, James Livsey, Scott Holan (R&M)

J. Imputation Modeling for Multivariate Categorical Characteristic Data in 2030 Census

Description: Nonresponse and administrative record enumeration in the Decennial Census led to item missingness for person and household characteristics. The 2020 Census used past census and administrative record data to directly assign characteristics when missing. For the 2030 Census, staff are researching statistical imputation models for multiple categorical variables. The broad goal of this project is to study how to make better use of statistical modeling, in conjunction with administrative records, to enhance previously implemented procedures for characteristic imputation.

Highlights: During FY 2023, staff had regular conversations with Decennial Statistical Studies Division colleagues to review and discuss results from group quarters multiple imputation for categorical characteristic information. Staff provided advice and suggestions for multiple chained equation methods in R, and implementation of edit and imputation methods in Python.

Staff: Darcy Steeg Morris (x33989), Joseph Kang, Joseph Schafer (R&M)

K. Group Quarters Count Expectation Modeling to Ensure Data Quality

Description: The project objective is to develop a procedure which supplies expected counts for each group quarter (GQ) prior to 2030 Census production. Doing so will allow managers of the GQ operations to know if reported GQ counts fall outside an expected range. This information will be helpful so problems can be remedied before and while GQ data collection is ongoing.

Highlights: During FY 2023, staff considered various models for predicting the number of occupants in college dormitories and in nursing homes, respectively. Various statistical diagnostics were considered for model comparison and selection. As well, staff considered various approaches for prediction and outlier detection in association with each of these scenarios and worked to identify commonalities among quarters with similar types and/or scale of outliers.

Staff: Kimberly Sellers (x39808), Andrew Keller (DSSD)

L. Mobile Questionnaire Assistance: Analysis and Simulation

Description: Mobile Questionnaire Assistance (MQA) is an outreach program where Census Bureau staff organize events - often in communities anticipated to have lower response rates - to encourage response to the census and assist with the process of responding. Such outreach is believed to reduce workload in advance of a Nonresponse Follow-up operation. This project studies the impact of MQA operations on response rates. Data recorded in the 2020 Census will be analyzed for evidence of this relationship via statistical modeling. Insight from data analysis will be used to consider a simulation framework which could aid future design of MQA operations.

Highlights: During FY 2023, staff participated in regular discussions on the MQA project including model development, future census tests, and other ongoing project activities. An improved formulation of the "proximity" metric was developed to express a tract's exposure to the MQA operation in space and time, with computationally intensive integrals computed independently of tuning parameters. An extension to the proximity metric was proposed to scale exposure with population density so that an MQA event taking place 20km away (for example) contributes less to exposure in a more densely populated region than it does in a more sparsely populated region. Staff explored two variations of statistical models to associate a tract's proximity to its response rate: a simpler setting relating overall census response rates to MQA proximity aggregated over the census operation, and a setting which associates response rates and MQA proximities observed over time during the census operation. Several variations of linear and nonlinear models were considered using an appropriate transformation of response rate as the outcome. Working with stakeholders, staff developed use cases for model-assisted placement, including evaluation of batches of placements - by locations and dates – which have been selected by planners, and identifying "hot spots" where additional placements may provide the most benefit. Staff maintained internal R packages to facilitate eventual dissemination of modeling work to stakeholders. Staff developed interactive graphical interfaces with R Shiny to demonstrate model-assisted placement on maps and to visualize curves of response rate as a function of MQA proximity. A manuscript describing the project is in progress.

Staff: Andrew Raim (x37894), Lisa Moore (DCMD), Doug Fletcher (DSSD)

M. Exploring the Association between Training and Field Performance in 2020 U.S. Census Address Canvassing

Description: The 2020 U.S. Decennial Census included an Address Canvassing field operation to update address

lists. Prior to beginning field work, field staff received training and completed a training assessment pertinent to the operation. An assessment score is used to determine staff's mastery of knowledge and skills required for performing the field work. The purpose of this investigation was to examine the association between assessment score and field performance.

Highlights: During FY 2023, staff constructed two binary field performance indicators, based on quality control field data at the housing unit level: (i) a "critical" error (which results in coverage error), and (ii) any error (which may result in coverage error or other non-critical errors). Staff conducted Cochran-Armitage trend tests to explore the trend between performance errors and the training assessment scores of field staff. The trend tests were repeated on four sub-domains of the assessment (knowledge, data entry, workflow, and navigation). The preliminary results revealed a significant trend of improved performance with an increase in assessment scores. The same was true for each of the sub-domains, with the knowledge domain showing the strongest trend.

A manuscript "Exploring the Association between Training and Field Performance in 2020 U.S. Census Address Canvassing" has been submitted to the *2023 Proceedings of the American Statistical Association.*

Staff: Thomas Mathew (x35337), Mathew Virgile (CBSM), Lin Wang (CBSM)

N. Assessments of the 2020 Nonresponse Follow-up Enumerator and Census Field Supervisor Training
Description: This investigation is a continuation of the previous analysis on the association between assessment score and field performance. The nonresponse follow-up enumerator final assessment contained 15 items. The research questions of interest are: (i) Does the field performance depend on the scores on each of the 15 items? (ii) How appropriate is the training assessment score weighting? (iii) How valid are the training assessment question items in measuring the trainee's acquisition of knowledge and skills? The enumerators included had training assessment scores of at least 70 and had at least one re-interview case.

Highlights: During FY 2023, a multifaceted approach was developed to carry out the study using existing operational data. We first conducted a comprehensive review of the NonResponse Follow-up (NRFU) operational process, and then developed primary and secondary field performance indicators based on available data. Multiple data sources were wrangled to construct appropriate data sets for analyses. A standard chi-square analysis showed the strong association between the field performance and the score on each of the 15 items. Effects sizes were calculated based on the Cochran's V statistic in order to order the questions in terms of their effect on field performance. The preliminary results showed a significant trend of improved performance, though small in effect size, with an increase in assessment scores. In order to explore the effect of the training assessment score weighting, the association between the raw scores and the field performance is under investigation.

Staff: Thomas Mathew (x35337), Mathew Virgile (CBSM), Lin Wang (CBSM)

O. Agreements for Advancing Record Linkage
Description: Motivated by the enhanced needs at the Census Bureau regarding the state-of-the-art methodology and algorithms for record linkage and entity resolution, three universities have been awarded priority one cooperative agreements: The University of Michigan, The University of Connecticut, and the University of Arkansas, Little Rock.

Highlights: During FY 2023, the universities focused on development and evaluation of methods and algorithms with illustrations on both simulated and real data. The universities gave progress presentations and received feedback from staff.

Staff: Rebecca Steorts (919-485-9415), Emanuel Ben-David, Dan Weinberg, Krista Park (CODS), Anup Mathur (CODS)

P. Continuous Count Study
Description: The Continuous Count Study has two main parts. The first is to produce an administrative records enumeration by leveraging the work being done on the Demographic Frame. The second is to produce alternative estimates to evaluate the quality and coverage of this enumeration. This will initially be done for specific target dates (April 1, 2020 and July 1, 2021) but the plan is to do both the administrative records enumeration and the quality and coverage evaluation on an ongoing basis.

Highlights: During FY 2023, staff began examining and running some basic crosstabulations on the 2020 and 2021 Demographic Files and the 2020 Census Edited File (CEF). Staff wrote a draft memorandum documenting some of the exploratory comparisons between the 2020 Demographic File and the CEF. Staff ran four simple imputation methods on the 2021 Demographic File and ran one of them on the 2020 Demographic File. The first two methods use only Demographic File Data. Missing sex and age group are imputed with a previous person hot deck with a geographic sort. Within-household missing race and Hispanic origin are imputed with a previous person hot deck (or next person, if necessary) within the household. The first method uses a previous person hot deck with a geographic sort for whole household missing race and whole household missing Hispanic origin, the

second method uses a previous person hot deck with a geographic sort where missing race is imputed using persons with the same Hispanic origin and vice versa. The third method replaces Demographic File missing data using CEF data for persons who match to the CEF, with age group calculated using CEF date of birth and the Demographic File reference date. The fourth method uses CEF data for all Demographic File persons who match to the CEF (regardless of whether any data is missing or not). For both the third and fourth methods, any remaining missing data is imputed the same way as in the second method. The fourth method was also run on the 2020 Demographic File, adjusting the age group calculation for the 2020 reference date. Staff documented the results of the last three imputation methods on the 2021 Demographic File in a draft memorandum. One key result is that using CEF data tends to decrease the proportion of White Hispanics and increase the proportion of Multiracial Hispanics. first method. Finally, staff began examining files containing various sets of results relevant to the Continuous Count Study in preparation for looking at comparisons between the different sets of results.

Staff: Michael Ikeda (x31756)

Q. Capture-Recapture Coverage Measurement using Administrative Records (Continuous Count Study)
Description: This project focuses on an investigation into the use of administrative records and ongoing sample surveys to produce population estimates on a continuous basis using dual or multiple system estimation methodology.

Highlights: During FY 2023, this work began and involves estimating the 2020 Census population using a dual-system or triple-system estimation (capture-recapture). The three systems are: the 2020 Census, the 2020 ACS, and the demographic frame. The latter is obtained from a variety of administrative records. We are utilizing Joe Schafer's CVAM package to fit a model that can estimate the "hidden cells," which consists of people who were not found in all three systems. Demographic information, such as Hispanic origin, must be imputed in these cases. A latent class model can then be fit to predict the true demographic class given previously imputed values for all three systems.

We will also utilize a two-system estimation using the Census and demographic frame for people at Master Address File IDs that are not valid for ACS.

Staff: Dan Weinberg (x38854), Tom Mule (DSSD)

R. Cohort Component Birth Modeling (Continuous Count Study)
Description. This project focuses on an investigation related to activities in the Cohort Components Study, a Continuous Count Study Group subgroup. The cohort components are of interest in the Population Estimates Program. The cohort-component method is derived from the demographic balancing equation:

Population Estimate = #Base Population + #Births - #Deaths + #Net In-Out Migrations

The main objective of the Cohort Components Study is to explore the use of administrative records to obtain counts of births, deaths, and net-in-out-migrations. Officially, we obtain birth and death data from the National Center for Health Statistics, which has a two-year lag. The two-year lag means that the most recent final data on births and deaths by geographic and demographic detail for each vintage of estimates refer to the calendar year two years before the vintage year. For example, the most current full-detail births and deaths data used in Vintage 2022 were from 2020. We, however, received record-level birth **ad**ministrative **rec**ords (ADREC) in several ways during those two years. The main objective of this project is to investigate whether ADREC can help improve the county birth estimation.

Highlights: During FY 2023, in a preliminary explanatory analysis, staff used the Numident (Social Security Administration), the Demographic Frame, and ACS-pop estimates to provide counts of births in 2020 by county in Minnesota.

Staff: Emanuel Ben-David (x37275), Tom Mule (DSSD), Eric Jensen (POP), Esta Miller (POP)

S. Evaluating and Improving the Low Self-Response Score (LRS) with the 2020 Census Data
Description: Following the 2020 Census, the Census Bureau has formed the *LRS Workgroup* to work on updating and improving the LRS. The Workgroup is made up of statisticians, demographers, and technical experts in census and survey operations from across the Census Bureau. The outline of the in-scope and out-of-scope activities and key deliverables for the Workgroup is as follows. In-scope activities include: (1) survey current users of LRS, determine the needs of these users; (2) evaluate the current LRS model, test if the Census and ACS response rates result in a similar LRS model, update the LRS model with 2020 response rates, develop a new LRS model and additional summary scores for the PDB. Out-of-scope activities include making significant revisions to the PDB or ROAM tool, using administrative records or other nonpublic data sources (subject to change depending on how formal privacy will impact the Census operational statistics). Key deliverables are: (1) summary report on stakeholders for the PDB and LRS (2) evaluation report on the current LRS measure (3) revised LRS measures (4) additional summary measures for the PDB and (5) documentation on revised LRS and other scores.

Highlights: During the first two quarters of FY 2023, we

specified four main sub-projects for improving the LRS modeling. In these sub-projects, we planned to explore and evaluate: (1) opportunities for incorporating additional geographic improvements to the LRS model, including spatial modeling and incorporating additional spatial layers; (2) alternative data sources for adding covariates to current prediction model of LRS; (3) local regressions for improving the current LRS model using state-based models and also estimating the margin of error for the LRS score; and (4) the utility of the LRS model when the data are noise-infused files from CEDDA for privacy protection. Since the third quarter of FY 2023, we have suspended the activities mainly due to some core members transitioning to other positions at the Census Bureau. The activities are expected to resume on the assigned tasks sometime in FY 2024.

Staff: Emanuel Ben-David (x37275), Joanna Fane Lineback (CBSM), Eric Jensen (POP), Luke Larsen (CBSM), Kathleen Kephart (CBSM), Heather King (SEHSD), Steven Scheid (DSSD), Fang Weng (CBSM)

T. Record Linkage Support for Decennial Census
Description: In preparation for the 2030 Census, the Decennial Statistical Studies Division (DSSD) must evaluate the previous decennial census matching methodology. DSSD refers to this project as "Project 80." This project will evaluate and determine the matching methodology and software used for the next Census. The methodologies that will be evaluated include: the order of blocking, the matching parameters and their matching weights, nickname standardization, match modeling, and evaluation of matching categories. More software packages will be evaluated and tested to help improve the identification of duplicates.

Highlights: During FY 2023, staff joined "Project 80" of Decennial Statistical Studies Divisions (DSSD) to provide support software to assist in a smoother deduplication and evaluation. The linux scripts manage software such as *BigMatch* to run on a parallel platform, saving computation time. Staff provided support for blocking and name standardization strategies. Staff successfully ported *BigMatch* to a jupyter platform. Jupyter which means (Julia, Python and R). It allows for efficient documentation of software. The new documented *BigMatch* code adapted to new platform easily and identify duplicates of 2020 Census data.

Staff: Ned Porter (x31798), Dan Weinberg

1.14 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 6385J70)

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

Highlights: During FY 2023, staff completed a manuscript discussing methodology for generating custom ACS estimates from a continuous-time model, and submitted the paper for publication.

Staff: Tucker McElroy (240-695-3610; R&M), Patrick Joyce

B. Visualizing Uncertainty in Comparisons and Rankings Based on ACS Data
Description: This project presents results from applying statistical methods which provide statements of how good the rankings are in the ACS Ranking Tables [See The Ranking Project: Methodology Development and Evaluation, Research Section under Project 0331000].

Staff: Tommy Wright (x31702), Adam Hall, Nathan Yau, Jerzy Wieczorek (Colby College)

C. Voting Rights Act, Section 203 Model Evaluation and Enhancements Towards 2021 Determinations
Description: Section 203 of the *Voting Rights Act (VRA)* mandates the Census Bureau to make estimates every five years relating to totals and proportions of citizenship, limited English proficiency and limited education among specified small subpopulations (voting-age persons in various race and ethnicity groups called Language Minority Groups [LMGs] for small areas such as counties or minor civil divisions MCDs). The Section 203 determinations result in the legally enforceable requirement that certain geographic political subdivisions must provide language assistance during elections for groups of citizens who are unable to speak or understand English adequately enough to participate in the electoral process. The research undertaken in this project consists of the development, assessment and estimation of regression-based small area models based on 5-year American Community Survey (ACS) data and the Decennial Census.

Highlights: During FY 2023, work on this project consisted of writing a journal paper on the Voting Rights Act Section 203(b) modeling as a Small Area Estimation effort. The paper was written, submitted, shepherded through revision, and accepted.

Staff: Eric Slud (x34991), Adam Hall, Mark Asiala (DSSD), Joseph Kang, Tommy Wright

D. Voting Rights Act (VRA) Section 203 Research Towards 2026 Determinations (also Decennial Project 6550J06)
Description: The *Voting Rights Act* of 1965 prohibits discrimination in voting. Section 203 of the *Voting Rights Act* mandates the Census Bureau to make estimates every five years relating to totals and proportions of citizenship, limited English proficiency

and limited education among specified small subpopulations (voting-age persons in various race and ethnicity groups called Language Minority Groups [LMGs] for small areas such as counties or minor civil divisions MCDs). The Section 203 determinations result in the legally enforceable requirement that certain geographic political subdivisions must provide language assistance during elections for groups of citizens who are unable to speak or understand English adequately enough to participate in the electoral process. The research undertaken in this project consists of the development, assessment, and estimation of regression-based small area models based on 5-year American Community Survey (ACS) data, the Decennial Census, and administrative records. These models will be used to produce more accurate estimates in small areas for the 2026 determinations.

Highlights: During FY 2023, staff accomplished the following: (1) documentation of the standard operating procedure that reproduced the VRA outcomes without any technical errors, (2) preparation of both the Numident (from Social Security Administration) and IRS predictor covariates, and (3) implementation of the Bayesian version of Multinomial Random Effects model, which was optimized to run efficiently.

Staff: Joseph Kang (x32467), Adam Hall, Xiaoyun Lu, Yathish Kolli (CODS), Amandeep Bajawa (CODS)

E. Assessing and Enhancing the ACS Experimental Weighting Approach Implemented in 2020 Data Products

Description: Census Bureau sample survey data exhibited unprecedented levels of missing data in 2020 because of data collection interruptions due to the COVID-19 pandemic. With administrative record linked data, Rothbaum and Bee (2021) documented differences in characteristics between ACS respondents and non-respondents, suggesting that nonresponse bias may affect estimates in the 2020 data. Experimental nonresponse weights were developed using a calibration technique (entropy balancing) based on demographic and administrative record (e.g., income) benchmarks (Rothbaum et al., 2021). The goal of this project is to study the experimental weighting methodology to assess its performance in simulated data scenarios, and to compare it to alternative nonresponse weighting techniques (e.g., inverse propensity weighting-IPW). In developing a deeper understanding of the experimental weighting, staff may also study improvements on the experimental weighting such as accounting for benchmarking to totals estimated from the administrative data and benchmark variable selection.

Highlights: During FY 2023, staff continued regular conversations with colleagues in the Decennial Statistical Studies Division, Center for Economic Studies, and Social, Economic, and Housing Statistics Division to understand ACS data collection processes including CAPI subsampling, clarify structure and limitations of ACS process and administrative records (AR) linked data, and variations of sample survey selection weights. Staff worked with a research dataset on response status for and administrative records for 2018-2021 ACS sampled units. With this dataset, staff looked at a variety of basic statistics such as geographic and monthly comparison of response rates and administrative record match rates.

Staff continued development and comparison of logistic regression, lasso, random forest and generalized boosting models for modeling response propensity. Staff continued to refine the models based on ACS process and theoretical considerations understood through regular conversations with ACS experts. These model developments led to response probability estimates that are used in an alternative nonresponse adjustment approach that (1) adjusts weights with the inverse of the response probability estimate to balance respondent and nonrespondent characteristics and then (2) calibrates to benchmarks. Staff developed metrics to use for model comparison to include practical considerations as well as traditional checks such as comparison of weighted means of respondents to the entire sample for a large set of characteristics. Results were consistently assessed and discussed to continue to further refine the methodology. Staff also researched and documented information about incorporating survey weights in lasso regularized logistic regression models.

Staff prepared and presented results on response propensity model comparisons for the 2018-2021 ACS data internally and in an invited session at the Joint Statistical Meetings on "Using Machine Learning & Data Science Methods to Improve Model Assisted Estimation & Survey Design." Based on these model comparison results, staff ran IPW weighting using logistic regression and GBM for all states and all four years to be used to produce final weights and thus experimental ACS estimates. Staff began the visual evaluation and comparison of ACS estimates at varying levels of geography through an R Shiny app. Staff engaged in discussions about objective measures of evaluation, particularly with respect to how to account for the vast amount of ACS outcomes that could be assessed.

Also, as part of the research on this project, staff [Eric Slud] developed theory of two-stage weighting adjustment for possible application to ACS weighting. The method involves a first stage of model-assisted Inverse Probability Weighting, where the first stage uses base-weights and involves fitting a model like logistic regression or solving a calibration-style estimating equation, respectively using either covariate data on all sampled units or externally known target population-proportions for covariates. At a second stage, the first-stage adjusted weights are used in place of base-weights,

and weights are adjusted via calibration (e.g., raking) to known frame-population totals. Theoretical justification of such a two-stage approach is given in a design-based framework, and variances of survey-weighted outcome-total estimators are calculated via linearization. This research forms the basis of the author's invited International Statistical Institute's 2023 Conference presentation, in which the method is illustrated on data from the 2020 Census Bureau Tracking Survey.

Staff: Darcy Steeg Morris (x33989), Joseph Kang, Patrick Joyce, Isaac Dompreh, Eric Slud, Shane Lubold, Tommy Wright

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

A. Research on Biases in Successive Difference Replication Variance Estimation in Very Small Domains

Description: In various small-area estimation contexts at the Census Bureau, current methods rely on the design-based sample survey estimates of variance for survey-weighted totals, and in several major sample surveys including the American Community Survey (ACS) and Current Population Survey. These variance estimates are made using Successive Difference Replication (SDR). One important application of such variance estimates based on ACS is the *Voting Rights Act (VRA)* small-area estimation project supporting the Census Bureau determinations of jurisdictions mandated under *VRA* Section 203(b) to provide voting language assistance. The current research project is a simulation-based study of the degree of SDR variance-estimation bias seen in domains of various sizes.

Highlights: During FY 2023, the simulation code on this project was refined and extensively tested, and theoretical formulas were developed and written up to express in terms of design and superpopulation parameters the bias and relative bias in SDR variance estimation for sample surveys for stratified SRS designs with small sampling fraction. Using the formula and simulation code, several examples have been developed of superpopulations in which SDR biases on small domains would be larger than has previously been documented. These results will be written up into a technical report and paper in FY 2024.

Staff: Eric Slud (x34991), Tim Trudell (DSMD)

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

A. Data Integration

Description: This Research looks at linking Current Population Survey (CPS) with other data sources for two purposes: (1) To ensure the public use microdata files cannot be used to identify participants in the CPS and (2) To see if alternative data sources (for example, ACS and administrative data) can be used to improve or independently produce CPS statistics. Staff will proceed by gaining deep knowledge and understanding of the Current Population Survey and the American Community Survey.

Highlights: During FY 2023, staff started a project to identify people who participated in the Current Population Survey by linking PUMS data with administrative lists which are available to the public. Currently data cleaning software is needed to eliminate un-identifiable records. Data from the administrative lists of the re-identification attack are to be classified as either in-scope or out-of-scope records. An out-of-scope record in this case, could not possibly be re-identified accurately due to missing data or other factors. Software is being developed to clean the data of all out-of-scope records. Further data cleaning is required when metadata between administrative lists do not agree with the CPS released data. Staff is developing software using machine learning classifiers to complete this phase. Staff have analyzed the metadata from the CPS.

Staff shared a plan for detecting curbstoning using machine learning methods. This involved records that had been flagged as false to be used as training data. The training data would be applied using machine learning algorithms, such as decision trees. This plan was presented to the Improving Quality for Enumeration and can be only used on Census Data.

Staff: Ned Porter (x31798), Emanuel Ben-David

1.17 SOCIAL, ECONOMIC, & HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165023)

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 sample surveys) in producing the desired estimates. Also included would be an evaluation of the techniques developed, along with documentation of the methodology.

Highlights: During FY 2023, staff has explored generalizing the Graphically Weighted Regression (GWR) extension to the Fay-Herriot model to a more general area-level small area model framework. Specifically, staff is targeting the Dirichlet-Multinomial shares model which is being developed for school district population and poverty estimates. Staff has also worked on deriving an estimation procedure that allows only a subset of the parameters to vary across observations while the remainder of the parameters are constant over all observations.

Also, staff (along with their SEHSD counterparts) have been assisting the Data Integration subteam of the Continuous Count Study project. The staff has ported over the implementation of the Multinomial Dirichlet Small Area model code to use for estimating the tract to county share of the population living in housing units. The results (and tract-level population predictions) of the modeling were presented to the full Continuous Count Study group in August 2023.

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

B. Assessing Constant Parameters across Areas in the SAIPE Models

Description: In the SAIPE production models, there is an assumption that the covariates have the same relationship with the outcome variable (number of school-age children in poverty) across all areas (state, county, school districts) and that the error variance is homogeneous.

There is great variability between the counties (and school districts) in terms of population size, racial composition, general economic statistics, etc. which may have interactions effects on subsets of the areas. Staff will develop methods to evaluate the assumption of a constant uniform relationship of the parameters across all areas for the SAIPE county (and eventual school district) poverty models.

Highlights: During FY 2023, staff members have been writing and revising a research report which documents the research done to bring the ideas of Geographically Weighted Regression (GWR) into the Fay-Herriot small area model framework. Staff have developed a graphical method to detect when parameters vary across areas compared to a singular model where parameters are the same for all areas. Staff is planning to use the GWR method to evaluate the SAIPE county production model. Several different dimensions for the weighting have been identified such as county population size and income-tax based child poverty rates. Staff is also still searching for other factors that may be useful to incorporate into the weight function.

Staff: Jerry Maples (x32873), Isaac Dompreh, Wes Basel (SEHSD)

C. 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.

Development of unit-level small area modeling strategies under informative sampling designs.

Highlights: During FY 2023, staff created an R package for efficient fitting of unit-level small area models under informative designs. This R package includes options for fitting to both continuous and discrete data, as well as to data which is clustered. A draft of a statistical software paper documenting this package is being written.

Staff: Ryan Janicki (x35725), Paul Parker, Scott Holan (R&M)

1.18 GENERAL ECONOMIC STATISTICAL SUPPORT (Economic Project 1183X01)

1.19 GENERAL ECONOMIC STATISTICAL PROGRAM MANAGEMENT (Economic Project 1183X90)

A. Use of Big Data for Retail Sales Estimates

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

Highlights: During FY 2023, staff worked with colleagues in the Economic Statistical Methods Division and the Research and Methodology Directorate to complete, submit, and re-submit (after receiving a revise and resubmit request) a research paper to a peer-reviewed journal on a hierarchical Bayesian mass imputation methodology for estimating state-level retail sales based on data from a third-party aggregator.

Staff: Darcy Steeg Morris (x33989), Stephen Kaputa (ESMD), Rebecca Hutchinson (EID), Jenny Thompson (ESMD), Tommy Wright

B. Seasonal Adjustment Support

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

Highlights: During FY 2023, staff provided seasonal adjustment and software support for users within and outside the Census Bureau, including the Federal Statistical Office of Germany, Union Bank of the Philippines, Bank of Israel, Central Bureau of Statistics in Israel, Instituto Nacional de Estadistica y Censos (Argentina), Bloomberg Economics, Vienna Institute of Demography, California Department of Tax and Fee Administration, Bureau of Labor Statistics, California Department of Finance, The Bank of Nova Scotia, Instituto Nacional de Estadistica y Censos (Argentina), Bureau of Labor Statistics, Sophia Group (India), i4f Patents and Technologies (Belgium), XAI Asset Management, Wake Forest University, Finatext, Florida State Legislature, and Australian Bureau of Statistics, as well as several private users. Staff continued to update the GitHub repository for Ecce Signum.

Staff: Tucker McElroy (240-695-3610; R&M), James Livsey, Osbert Pang, William Bell (R&M)

C. Seasonal Adjustment Software Development and Evaluation

Description: The goal of this project is a multi-platform computer program for seasonal adjustment, trend estimation, and calendar effect estimation that goes beyond the adjustment capabilities of the Census X-11 and Statistics Canada X-11-ARIMA programs, and provides more effective diagnostics. The goals for FY 2023 include: continuing to develop a version of the X-13ARIMA-SEATS program with accessible output and updated source code so that, when appropriate, the Economic Directorate can produce SEATS adjustments; and incorporating further improvements to the X-13ARIMA-SEATS user interface, output and documentation. In coordination and collaboration with the Time Series and Related Methods Staff of the Economic Statistical Methods Division (ESMD), staff will provide internal and/or external training in the use of X-13ARIMA-SEATS and the associated programs, such as X-13-Graph, when appropriate. Additionally, development efforts are focusing on the future software products that advance beyond current capabilities of X-13ARIMA-SEATS. This new product aims to handling sampling error, treatment of missing values, and multivariate analysis. This development is a joint effort with staff from the Center for Optimization & Data Science and the Economic Statistical Methods Division.

Highlights: During FY 2023, Build 59 of X-13ARIMA-SEATS was publicly released, and the internal testing of the subsequent build (Build 60) was initiated. We continued ongoing discussions focused on the SEATS output structure. Simultaneously, the development of a Python program to run X-13ARIMA-SEATS progressed. Including an emphasis on incorporating regComponent usability, while the GUI for the interface remained in a beta version. The Economic Directorate staff included an update and live demonstration of pyX13 by Time Series planning group members from ECON and CODS. User feedback led to the identification and rectification of an omission in documentation, addressing a threshold issue for selecting the Henderson trend filter in monthly series within the X13 manual, an improvement subsequently integrated into the manual.

Staff: James Livsey (x33517), Demetra Lytras (ESMD), Tucker McElroy (R&M), Osbert Pang, William Bell (R&M), Lijing Sun (CODS)

D. Research on Seasonal Time Series - Modeling and Adjustment Issues

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

Highlights: During FY 2023, staff continued work on several projects, including: (a) applied maximum entropy outlier framework to time series affected by Covid-19, extending the methodology to simultaneously handle extreme values and missing values. These extended methods were used to generate seasonal adjustments with

uncertainty based on the extreme value adjustment; (b) continued development and implementation of a principal components methodology for the Census IDEA index, extending the nowcasting framework (by researching an estimation method that is less sensitive to extreme values) and exploring a mixed frequency extension; (c) developed and applied benchmarking techniques to constituent variables of GDP, in order to remove residual seasonality while preserving aggregation relations. The reconciliation methodology maintains frequency and hierarchical aggregation relations, subject to seasonal adjustment adequacy; (d) further developed weekly time series models that take account of the non-integer period of seasonality, examining different kinds of fixed effects; (e) continued the study of multivariate model-based seasonal adjustment applied to manufacturing time series, and made comparisons to univariate adjustment; and (f) revised writing and simulations for a study of mean squared error of seasonal adjustments from differing frameworks.

Staff: Tucker McElroy (240-695-3610; R&M), James Livsey, Osbert Pang, William Bell (R&M)

E. Supporting Documentation and Software for Seasonal Adjustment

Description: The purpose of this project is to develop supplementary documentation and utilities for all software related to seasonal adjustment and signal extraction at the Census Bureau. Staff members document X-13ARIMA-SEATS that enable both inexperienced seasonal adjustors and experts to use the program as effectively as their backgrounds permit. Ecce Signum, the Census Bureau's R package for multivariate signal extraction, documentation is being developed for submission to the *Journal of Statistical Software*. This fiscal year's goals include improving the X-13ARIMA-SEATS documentation, exploring the use of R packages that interface with X-13ARIMA-SEATS.

Highlights: During FY 2023, changes to the reference manual were synchronized with the public release of Build 60 of X-13ARIMA-SEATS. This involved updating the SEATS spec to incorporate tables providing standard errors, alongside ongoing revisions to the composite spec, refining descriptions of output tables and their usage. We improved the X-13 manual concerning the threshold for selecting the Henderson trend filter, explicitly stated for monthly series but overlooked for quarterly series. This discrepancy was diligently addressed, and the programmed values were subsequently added to the manual. Further adjustments were implemented in the composite spec, specifically fixing the value type for the indrndsachanges table, and correcting the description of the indsadjround value from a percent change to the accurate output of "ROUNDED INDIRECT SEASONALLY ADJUSTED SERIES." Additionally, parms arguments were introduced to the 'rarely used' section of the estimate spec.

Staff: James Livsey (x33517), Tucker McElroy (R&M), Osbert Pang, William R. Bell (R&M)

F. Exploring New Seasonal Adjustment and Signal Extraction Methods

Description: As data become available at higher frequencies and lower levels of disaggregation, it is prudent to explore modern signal extraction techniques. This work investigates two model-based signal extraction methods with applications to the U.S. Census Bureau's M3 survey: signal extraction in ARIMA time series (SEATS) and multivariate signal extraction with latent component models. We focus on practical implications of using these methods in production, focusing on revisions and computation complexity.

Highlights: During FY 2023, staff engaged in a comprehensive exploration of seasonal adjustment methodologies by developing and implementing custom code to assess the performance of multivariate seasonal adjustments in comparison to univariate adjustments. Preliminary findings indicate that, when accurately specified, the multivariate approach demonstrates superior performance over its univariate counterpart. Transitioning into the empirical phase of the study, the team initiated the modeling of M3 series and crafted code to directly extract the latest time series data from the Census Bureau's API. Additionally, recognizing the significance of signal extraction, our team embarked on an empirical study to investigate Granger causality, further enhancing our understanding of the intricate dynamics within the data analysis process.

Staff: James Livsey (x33517), Colt Viehdorfer (ESMD), Osbert Pang

G. Production and Dissemination of Economic Indicators

Description: In this project, we investigate potential improvements to the production and dissemination of economic indicators.

Highlights: During FY 2023, staff continued to work with collaborators at the University of Michigan to recalculate CUPI and SUPI indices, and to generate bootstrap replications to estimate the uncertainty associated with these indices.

Staff: Adam Hall (x32936)

H. Small Area Estimation for the Annual Integrated Economic Survey

Description: The Annual Integrated Economic Survey (AIES) is a re-engineered sample survey designed to integrate and replace seven existing annual business sample surveys into a streamlined single survey

instrument. The goal of this project is to develop and implement small area estimation methodology to produce state level estimates for all AIES core items by three-digit North American Industry Classification System (NAICS3) groups.

Highlights: During FY 2023, staff worked to create a working dataset to explore Fay-Herriot style small area models for sales and employment for the retail sales sector. From initial work on the AIES, a sample based on the sample survey design was drawn from the 2017 Economic Census. Staff identified predictor variables that would have been available if the survey were actually taken in 2017 and is currently building a research dataset. One of the datasets identified was the Business Register.

Staff created a sample for the manufacturing sales sector based on the AIES sample survey design had it been drawn from the 2017 Economic Census. Staff cleaned and transformed historical business registry data for use as predictor variables. Staff identified an issue with implausible outliers in the business registry data and are working on a procedure to flag such outliers. Staff are currently exploring different covariate models for incorporating the business registry data into Fay-Herriot style small area models to predict sales and employment for the retail and manufacturing sales sectors.

Also, staff from our Center for Statistical Research and Methodology presented a set of small area methodology tutorials to help train the Economic Statistical Methods Division staff.

Staff: Jerry Maples (x32873), Serge Aleshin-Guendel, Gauri Datta, Ryan Janicki, Jenny Thompson (ESMD), Stephen Kaputa (ESMD), Aja Maison (EMD)

1.20 PROGRAM DIVISION OVERHEAD (Census Bureau Project 0331000)

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

Staff: Tommy Wright (x31702), Joseph Engmark, Michael Hawkins, Eric Slud, Kelly Taylor

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

Highlights: During FY 2023, we continued to maintain and improve the Integrated Research Environment (IRE) and the Cloud Research Environment (CRE). We updated the IRE hosts to RedHat Enterprise Linux 8 (RHEL 8), and updated Stata to version 18. We introduced a new editor – VS Code and tested a GIS package called GRASS to replace GeoDA (which no longer works on RHEL 8). We added a new "cache priming" feature to address timeout issues during login. We explored NICE DCV as a potential replacement for NoMachine Terminal Server and added many new Stata and R packages. On the Windows side we worked with the R User's Group to test its latest release of R, RStudio, and RTools for laptops and VDI, and assisted with the transition of switching from the Computer Services Division maintained CRAN mirror and Anaconda package repositories to Applications Development & Services Division's Sonatype (Nexus) repositories. Finally, we participated in the Internal Revenue Service's Safeguard Review.

Staff: Chad Russell (x33215)

1.21 NATIONAL CANCER INSTITUTE (Census Bureau Project 9401021)

A. Modeling Tobacco Use Outcomes with Data from Tobacco Use Supplement - Current Population Survey
Description: During the first and second quarters of FY 2017, staff started a new project using Current Population Survey (CPS) files from the 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. The Tobacco Use Supplement to the Current Population Survey (TUS-CPS) is a National Cancer Institute (NCI) sponsored sample survey of tobacco use that has been administered as part of the U.S. Census Bureau's Current Population Survey every two to four years since 1992. The TUS-CPS is designed to produce reliable estimates at the national and state levels. However, policy makers, cancer control planners, and researchers often need county level data for tobacco related measures to better evaluate tobacco control programs, monitor progress in the control of tobacco use, and conduct tobacco-related research. We were asked to help provide the county level data for NCI.

Highlights: During FY 2023, staff performed weighted regression analysis and Bayesian hierarchical models to produce county-level direct survey-based estimates of thirteen tobacco smoking outcomes. County-level designed-based estimates for these tobacco outcomes were calculated for 3,134 counties across the country. Additionally, model-based estimates for thirteen tobacco smoking outcomes were produced for 3,134 counties using 2018/2019 TUS-CPS files. Bayesian Hierarchical modeling through a Markov Chain Monte Carlo simulation was used to produce the final model-based county-level estimates for these thirteen tobacco outcomes.

Additionally, staff developed four small area estimation (SAE) models with the synthetic data approach to fit logistic regression models for six tobacco smoking outcomes with small sample sizes/proportions. The SAE models developed at the distributional levels include:
(a) Area-level probability small area model, this is an extension to the commonly used area level Fay-Herriot model with nested probability model.
(b) Arcsine-scale area-level small area model for proportion data to stabilize the variance. The arcsine transformation was used to stabilize the models with small sample size and high variance.
(c) Area-level binomial-logit small area model.
(d) Beta-logistic model with unknown sampling variance.

Staff: Isaac Dompreh (x36801), Benmei Liu (NCI)

2. RESEARCH

2.1 GENERAL RESEARCH AND SUPPORT (Census Bureau Project 0331000)

Missing Data & Observational Data Modeling

Motivation: Missing data problems are endemic to the conduct of statistical experiments and data collection operations. The investigators almost never observe all the outcomes they had set to record. When dealing with sample surveys or censuses this means that individuals or entities in the survey omit to respond or give only part of the information they are being asked to provide. Even if a response is obtained the information provided may be logically inconsistent, which is tantamount to missing. Agencies need to compensate for these types of missing data to compute official statistics. As data collection becomes more expensive and response rates decrease, observational data sources such as administrative records and commercial data provide a potential effective way forward. Statistical modeling techniques are useful for identifying observational units and/or planned questions that have quality alternative source data. In such units, sample survey or census responses can be supplemented or replaced with information obtained from quality observational data rather than traditional data collection. All these missing data problems and associated techniques involve statistical modeling along with subject matter experience.

Research Problems:
- Simultaneous imputation of multiple sample survey variables to maintain joint properties, related to methods of evaluation of model-based imputation methods.
- Integrating editing and imputation of sample survey and census responses via Bayesian multiple imputation and synthetic data methods.
- Nonresponse adjustment and imputation using administrative records, based on propensity and/or multiple imputation models.
- Development of joint modeling and imputation of categorical variables using log-linear models for (sometimes sparse) contingency tables.
- Statistical modeling (e.g., latent class models) for combining sample survey, census, or alternative source data.
- Statistical techniques (e.g., classification methods, multiple imputation models) for using alternative data sources to augment or replace actual data collection.

Potential Applications:
Research on missing data leads to improved overall data quality and estimate 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 for nonresponse and statistical modeling for using administrative records or alternative source data is important to supplement actual data collection in situations where collection is prohibitively expensive in Decennial, Economic and Demographic areas.

A. Data Editing, Imputation, and Weighting for Nonresponse
Description: This project covers development for statistical data editing, imputation, and weighting methods to compensate for nonresponse. Our staff provides advice, develops computer programs in support of demographic and economic projects, implements prototype production systems, and investigates edit, imputation and weighting methods theoretically and practically. Principled methods allow us to produce efficient and accurate estimates and higher quality microdata for analyses.

Highlights: During FY 2023, staff worked towards a deeper understanding of traditional missing data methodologies such as imputation and nonresponse weighting, with the purpose of re-thinking these methods in light of large-scale and biased missingness from decreasing response rates, data collection interruptions and survey design. To this end, staff is studying historic and recent literature on inverse probability weighting (IPW) and calibration techniques for nonresponse.

Staff continues hands-on learning of such methods through research motivated by studying the ACS experimental weighting procedure for 2020 ACS data products and working groups sharing ideas for alternate methodologies. Motivated by this application, staff has been researching alternative weighting methodologies and implemented a study of a two-stage (inverse probability followed by calibration) approach for adjusting for dynamic and large magnitude unit missing data. This study developed knowledge in response propensity modeling with machine learning techniques, as well as study of how machine learning techniques can be incorporated into traditional survey processing. Properties and performance of the two-stage approach were assessed based on a well-established simulation study from the causal inference literature. This work compares this IPW/calibration with GBM response modeling to traditional outcome-based imputation modeling, and IPW alone using logistic regression or GBM. This work was presented and discussed internally and at the Joint Statistical Meetings and published in the *Wiley Interdisciplinary Reviews (WIRE): Computational Statistics*. Staff began furthering this work with empirical studies of estimation bias comparisons for imputation with machine learning techniques versus IPW. This study is based on publicly available ACS data and aims to

expand on IPW only comparisons in the WIRE publication.

Staff is also building knowledge on modeling to jointly edit and impute for multivariate categorical variables. Motivated by the related project for 2030 Census characteristic imputation, staff read literature on simultaneous edit and imputation via Bayesian hierarchical models, provided feedback on and helped develop code for preliminary implementation of the models on 2020 Census data.

Staff: Darcy Steeg Morris (x33989), Joseph Kang, Isaac Dompreh, Shane Lubold, Yves Thibaudeau, Jun Shao

B. Imputation and Modeling Using Observational/Alternative Data Sources

Description: This project covers development of statistical methods and models for using alternative source data to supplement and/or replace traditional field data collection. Alternative source data includes administrative records – data collected by governmental agencies in the course of administering a program or service – as well as commercial third party data. Such data often contains a wealth of information relevant to sample surveys and censuses, but suffers from bias concerns related to, for example, coverage and timeliness. Imputation, classification, and general statistical modeling techniques can be useful for extracting good information from potentially biased data.

Highlights: During FY 2023, staff worked with Economic Statistical Methods Division (ESMD) staff to edit and submit a research paper describing the hierarchical Bayesian mass imputation model methodology for estimating state-level retail sales based on data from a third-party aggregator. An imputation model is built using the third-party data and applied to obtain imputations for all establishments in the survey frame. The imputed dataset is then used as input for the Monthly State Retail Sales (MSRS) – a more geographically granular and timely estimate than the produced Monthly Retail Trade Survey (MRTS). The purpose of the paper is to illustrate the usefulness of Bayesian multiple imputation hierarchical models (and ease of fitting with off-the-shelf software) for official estimates about the economy using third party data.

Staff is also assessing and studying the use of administrative record information in traditional imputation and nonresponse weighting methodologies. The projects described in part A introduce the novelty and availability of administrative record data to serve as both predictors in imputation and response propensity models, as well as to serve as benchmarks in calibration approaches. As part of those research projects, staff is interested in developing procedures for proper use and proper uncertainty quantification when using alternative data sources in missing data models.

Staff also began conversations and developing ideas about the use of alternate data in low response surveys with potentially large scale nonresponse bias. Staff is working with colleagues across the Census Bureau to consider larger-than-usual nonresponse in traditional surveys such as the ACS, as well as quick-turnaround electronic surveys with low response rates. Staff began developing research questions regarding nonresponse bias comparisons for alternate source data variables with varying levels of nonresponse, and the potential use of network questions to alleviate nonresponse bias through obtaining information about a respondent's heterogeneous network.

Staff: Darcy Steeg Morris (x33989), Joseph Kang, Isaac Dompreh, Shane Lubold, Yves Thibaudeau

C. Missing Data: Phase I-Jun Shao Lectures/Phase II-Problems, Applications, & Software

Description: This project provides a series of lectures as a step towards development of a small community inside the Census Bureau with deep knowledge in statistical methods for compensating for missing data. The lectures are to be offered by Jun Shao under the title "Statistical Methods for Handling Incomplete Data." As these lectures concluded, the participants decided to host a series of presentations under the title "Missing Data: Problems, Applications, & Software."

Highlights: During FY 2023, the last presentation in a series of presentations under the title "Missing Data; Problems, Applications, & Software Presentations/Lectures" was given by Mark Asiala (October 21, 2022, some recent work on the American Community Survey). This project has been completed.

Staff: Tommy Wright (x31702), Darcy Steeg Morris, Jun Shao

Record Linkage & Machine Learning

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

and reduced/eliminated re-identification risk. Easy means of constructing synthetic make it straightforward to pass files among groups.

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

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

A. Regression with Sparsely Mismatched Data
Description: Statistical analysis with linked data may suffer from an additional source of non-sampling error that is due to linkage error. For example, when predictive models are of interest, in the linkage process, the response variable and the predictors may be mismatched or systematically excluded from the sample. In this research, we focus on the cases where responses reside in one file and predictors reside in another file. These variables are then paired up using an error-prone record linkage process. We nevertheless assume that only a small fraction of these pairs is mismatched. The goal of the research is then to develop efficient methodologies for adjusting the statistical analyses for bias or inconsistency introduced by linkage error.

Highlights: During FY 2023, we submitted a new manuscript titled "A General Framework for Regression with Mismatched Data Based on Mixture Modeling," in which we presented a general framework to enable valid post-linkage inference in the secondary analysis setting, in which only the linked file is given. The proposed framework covers various statistical models and can flexibly incorporate additional information about the underlying record linkage process. Specifically, we proposed a mixture model for pairs of linked records whose two components reflect distributions conditional on match status, i.e., correct match or mismatch. Regarding inference, we developed a method based on composite likelihood and the EM algorithm and an extension towards a fully Bayesian approach. Extensive simulations and several case studies involving contemporary record linkage applications corroborate the effectiveness of our framework. We submitted two manuscripts for a novel methodology for improving applications of modern predictive modeling tools to linked data sets subject to mismatch error. In these two manuscripts, we proposed a few adjustment methods for ensemble methods of bagging trees and random forests. We evaluated the performance of the proposed adjustment methods in an application involving social media. Specifically, for the application, we focused on observed Twitter activity measures and predicted socio-demographic features of Twitter users to accurately predict linked measures of political ideology that were collected in a designed survey, where respondents were asked for consent to link any Twitter activity data to their survey responses (exactly, based on Twitter handles). We showed that the proposed methodology implemented in R essentially recovers results that would have been seen with the actual data. We also prepared a manuscript for extending the general framework for post-linkage data analysis to the Cox model with survival analysis that concerns censored data.

Staff: Emanuel Ben-David (x37275), Guoqing Diao (GWU), Priyanjali Bukke (GMU), Martin Slawski (GMU), Brady West (UMICH-Ann Arbor)

B. Comparison of Entity Resolution Methods
Description: Work is underway on comparing Bayesian entity resolution methods and probabilistic entity resolution methods recently proposed in the literature that have open source software. Methods under consideration are those proposed by Marchant et al. (2021), Sadinle (2018), and Edmorando et al. (2018).

Highlights: During FY 2023, we worked on revisions of experimental analysis and comparisons, and we do not anticipate coming back to this project.

Staff: Rebecca C. Steorts (919-485-9415)

C. Almost All of Entity Resolution
Description: Whether the goal is to estimate the number of people that live in a congressional district, to estimate the number of individuals that have died in an armed conflict, or to disambiguate individual authors using bibliographic data, all these applications have a common theme - integrating information from multiple sources. Before such questions can be answered, databases must be cleaned and integrated in a systematic and accurate way, commonly known as record linkage, de-duplication, or entity resolution. In an article, we review motivational applications and seminal papers that have led to the growth of this area. Specifically, we review the foundational work that began in the 1940's and 50's that have led to modern probabilistic record linkage. We review clustering approaches to entity resolution, semi- and fully supervised methods, and canonicalization,

which are being used throughout industry and academia in applications such as human rights, official statistics, medicine, citation networks, among others. Finally, we discuss current research topics of practical importance.

Highlights: During FY 2023, staff published a review paper on entity resolution, and the project is complete.

Staff: Rebecca C. Steorts (919-485-9415)

D. Improvement to the E-M Algorithm for Record Linkage

Description: In record linkage, the E-M Algorithm is used to fit the matching parameters, such as the m and u probabilities which are used to calculate the match weights. A high match weight near 1 suggests a record pair is likely to be a match, whereas a low match weight near 0 suggests a record pair is likely to be a nonmatch. When we use the E-M Algorithm, we might realize convergence of these probabilities (i.e., parameters) to 0 or 1. Realizing such extreme values as 0 or 1 can create over confidence in assigning a match status. This project focuses on developing new methods to prevent the convergence to the values 0 or 1.

Highlights: During FY 2023, staff implemented new software to unduplicate the Census Unedited File (CUF). The software is designed to implement the Fellegi-Sunter methodology when the counts of the observed comparison patterns between records include a large proportion of 0 counts. In those situations, the traditional E-M Algorithm (Winkler, 1988; Sadinle & Fienberg, 2013) may fail to converge and the parameters are not estimable -the MLE does not exist. Staff developed software that operates orthogonal projections to contract the original parameter space of a log-linear model. The contracted parameter space has the same number of degrees of freedom as the EMLE (Fienberg & Rinaldo, 2012), which contains the information surviving a failed attempt to maximize the likelihood. Staff wrote R programs implementing algebraic-geometry algorithms to derive the contracted parameter spaces for the new likelihood maximization. Staff compared the application of this approach to cvam (Schafer 2021) a package implementing a regularization approach custom-made for latent class modeling. Cvam can produce pseudo-MLE with the same number of degrees of freedom as the nominal parameter space by introducing prior information, which can also be construed as a regularization. Staff conducted simulations to compare the two approaches. One result from the simulations was that when the design matrix was reduced from 3 parameters to 2 parameters convergence to 0 or 1 was completely avoided while estimating the true parameters reasonably well. Staff presented the research at the 2023 Joint Statistical Meetings and submitted a paper to the *2023 Proceedings of the American Statistical Association*. There are plans to submit to a journal.

Staff: Daniel Weinberg (x38854), Yves Thibaudeau

E. Recent Advances in Data Integration

Description: The availability of both sample survey and non-survey data sources, such as administrative data, social media data, and digital trace data, has grown rapidly over the past decade. With this expansion in data, the statistical, methodological, computational, and ethical challenges around integrating multiple data sources have also grown.

Highlights: During FY 2023, staff served as co-editor for a special issue of the *Journal of Survey Statistics and Methodology*, *Volume 11, Issue 3*, June 2023, https://doi.org/10.1093/jssam/smad009 during Q1/Q2. The project is now complete.

Staff: Rebecca C. Steorts (919-485-9415), Joseph W. Sakshaug (University of Michigan)

F. A Primer on the Data Cleaning Pipeline

Description: Recently, the statistical and methodological questions around data integration, or rather merging multiple data sources, has grown. Specifically, the science of the "data cleaning pipeline" contains four stages that allow an analyst to perform down stream tasks, predictive analyses, or statistical analyses on "cleaned data." This article provides a review of this emerging field, introducing technical terminology and commonly used methods.

Highlights: During FY 2023, staff wrote a review article on the data cleaning pipeline that was published in a special issue of the *Journal of Survey Statistics and Methodology, Volume 11, Issue 3*, June 2023, Pages 553-568, https://doi.org/10.1093/jssam/smad0017. The project is now complete.

Staff: Rebecca C. Steorts (919-485-9415), Anup Mather (CODS), Krista Park (CODS)

G. Monitoring Convergence Diagnostics for Entity Resolution

Description: The purpose of this project was to review convergence diagnostics within the Bayesian record linkage community, propose novel ones, and illustrate these using open source software.

Highlights: During FY 2023, staff reviewed the literature on convergence diagnostics within the Bayesian record linkage community, highlighting issues with mixing due to the high-dimensional nature of the parameter space.

Also, staff proposed new convergence diagnostics that are more appropriate and practical for illustrating a Markov chain Monte Carlo sampler that fails to

converge. Furthermore, staff illustrated this using methodology and experimental results and provided open source software.

Staff: Rebecca C. Steorts (919-485-9415), Serge Aleshin-Guendel

H. On Computing the Jaro Similarity Between Two Strings
Description: The point of this project was to propose a scalable Jaro distance metric with open source code.

Highlights: During FY 2023, staff reviewed common distance metrics used in record linkage. Then they proposed a linear-time algorithm for the Jaro distance metric, illustrating its performance on both real and simulated data. open source software was provided.

Staff: Rebecca C. Steorts (919-485-9415), Anup Mathur (CODS), Krista Park (CODS), Ken Hasse (R&M), Daniel Weinberg, Joyanta Basak (University of Connecticut), Ahmed Soliman (University of Connecticut), Nachet Deo (University of Connecticut), S. Sahni (University of Florida), Raj Sanguthevar (University of Connecticut)

I. Novel Blocking Techniques and Distance Metrics for Record Linkage
Description: The point of this project was to propose novel blocking methods and distance metrics for record linkage.

Highlights: During FY 2023, staff reviewed common blocking methods and distance metrics in the literature. Staff proposed novel blocking and distance methods and ran experiments on both real and simulated data sets. A related paper was accepted for publication.

Staff: Rebecca C. Steorts (919-485-9415), Daniel Weinberg, Nachet Deo (University of Connecticut), Raj Sanguthevar (University of Connecticut), Joyanta Basak (University of Connecticut), Ahmed Soliman (University of Connecticut)

J. Variational Beta Linkage
Description: The purpose of this project is to scale bipartite record linkage to hundreds of thousands of records in minutes.

Highlights: During FY 2023, staff proposed a scalable bipartite record linkage algorithm using a variation approximation for the first time, providing all derivations. Staff ran experimental results and wrote a research draft paper that is in progress.

Staff: Rebecca C. Steorts (919-485-9415), Serge Aleshin-Guendel, Brian Kundinger (Duke University)

K. Predicting Survey/Census Response Rates
Description: This study is motivated by the U.S. Census Bureau's well-known ROAM application, which uses a linear regression model trained on the U.S. Census Planning Database to identify low self-response tracts. A crowdsourcing competition around ten years ago revealed that machine learning methods based on ensembles of regression trees led to the best performance in predicting survey self-response rates; however, the corresponding models could not be adopted for the intended application due to their black-box nature. The proposed model in this work is a nonparametric additive model with a small number of main and pairwise interaction effects using 0-1-based penalization. From a methodological viewpoint, both computational and statistical aspects of the proposed estimator and its variants that incorporate strong hierarchical interactions are studied. The proposed algorithms available (open source on Git Hub) extend the computational frontiers of existing algorithms for sparse additive models to handle datasets relevant to the application we consider. The findings from the model on the U.S. Census Planning Database show that in addition to being useful from an interpretability standpoint, the proposed models also lead to predictions that appear to be better than popular black-box machine learning methods based on gradient boosting and feedforward neural networks, suggesting that it is possible to have models that have the best of both worlds: good model accuracy and interpretability.

Highlights: During FY 2023, a paper on this topic was submitted to the *Annals of Applied Statistics* and is tentatively accepted with revision.

Staff: Emanuel Ben-David (x39275), Shabil Ibrahim (MIT), Rahul Mazumder (MIT), Peter Radchenko (University of Sydney, Australia)

Sampling Estimation & Survey Inference

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

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

influential values in retail trade survey, and c) surveys of government employment.

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

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

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

A. Comparison of Probability (RDD) and Nonprobability in a Census Tracking System
Description: As part of a Decennial Census Evaluation project, the Census Bureau conducted a Tracking Sample Survey (through a contractor Young & Rubicam) on attitudes to the decennial census and their relationship to completing the census questionnaire. The sample survey was conducted in a probability-sampling (RDD telephone survey) and nonprobability web-panel mode, from September 2019 through June 2020. A secondary, methodological goal of the Tracking Survey data collection was to compare the effectiveness of the two modes, the RDD telephone survey versus the nonprobability sample. Staff in our Center for Statistical Research & Methodology were brought onto the project in early 2020, to help in evaluating and possibly improving the post-stratification weighting adjustment of these two data samples.

Highlights: During FY 2023, research done on this topic related to the Tracking Survey was the culmination of previous years' efforts, consisting of writing of final reports and a few additional analyses supporting the final reports on the Tracking Survey as part of a Decennial evaluation project. The written pieces related to documentation of staff's techniques for weighting the Telephone and Web surveys and comparing their results. In addition, staff provided re-weighted estimates and standard errors for a daily time-series and 7-day moving average of estimates from both surveys: these data were incorporated into other parts of Center for Behavioral Science Method's final reported analyses of survey trends.

In addition to the research specifically about the Tracking Survey, staff extended previous theoretical research on asymptotic theory (consistency and variance estimation) design for design-based calibration weighting. The extension was to two-stage weighting in which the first stage consists of model-based or model-assisted estimation of Inverse-Probability base weights, and the second stage was ordinary calibration or raking to externally sourced demographic/geographic population totals. This research was reported in an invited International Statistical Institute talk.

Staff: Eric Slud (x34991), Darcy Steeg Morris, Jennifer Hunter Childs (CBSM), Casey Eggleston (CBSM), Jon Krosnick (CBSM), Yazmin Trejo (CBSM), Shaun Genter (CBSM)

B. The Ranking Project: Methodology Development and Evaluation

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

Highlights: During FY 2023, staff added 1-Year American Community Survey (ACS) data for 2021 to the comparisons of each state with the other states and for the estimated overall rankings of the states and a joint confidence region for 88+ different American ACS topics. The 2 updated visualizations now provide ACS data for the years 2018, 2019, and 2021 as parts of The Ranking Project on the Center for Statistical research & Methodology's Internet site under "Statistical Research." See the three links The Ranking Project; Comparisons of A State with Each Other State; and Estimated Rankings of All States.

Staff: Tommy Wright (x31702), Adam Hall, Nathan Yau, Jerzy Wieczorek (Colby College)

C. Sampling and Apportionment

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

Sample Allocation

Highlights: During FY 2023, staff completed the draft of a paper which considers the new topic of "tightness" of a joint confidence region for an estimated overall ranking. He further considers things that determine the level of tightness including the introduction of a mathematical framework for optimally allocation of an overall sample size for the K = 51 populations to minimize an objective function subject to fixed n.

Staff: Tommy Wright (x31702)

Apportionment

Highlights: During FY 2023, there was no significant progress on this project.

Staff: Tommy Wright (x31702)

D. Understanding Chao's Method of Unequal Probability Sampling

Description: Chao (1982) presents an elegant and simple to execute algorithm for the selection of a probability proportional to size sample without replacement. The algorithm's published presentation is short, has limited details, and may not be known to many practicing survey sampling statisticians. Starting with a listing of the units in a finite population of N units, and each unit having an associated measure of size, the first n units on the list are designated as being in the sample at stage n. Then the next unit on the list but not in the sample, is selected for inclusion in the sample with probability proportional to size relative to the first n+1 units. If this next unit is actually not selected, the sample at stage n+1 is the same as it was at stage n. On the other hand, and if the n+1 unit is actually selected, it enters the sample and one of the sample units at stage n is removed with a specified probability. At each stage the sample has n units. This process continues and ends at stage N with a probability proportional to size sample of n units relative to the overall population of N units!

Highlights: During FY 2023, staff completed and published a somewhat longer research report than Chao (1982) aiming to provide more details, clarity, and proofs

associated with Chao's Method, hoping to promote greater consideration of it. The title of the draft is "Understanding Chao's Method of Probability Proportional to Size Sampling."

Staff: Tommy Wright (x31702)

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 sample surveys designed for national estimates do not provide large enough samples to produce reliable direct estimates for small areas such as counties and even most states. The use of valid statistical models can provide small area estimates with greater precision, however bias due to an incorrect model or failure to account for informative sampling can result. Methods will be investigated to provide estimates for geographic areas or subpopulations when sample sizes from these domains are inadequate.

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

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

A. Bootstrap Mean Squared Error Estimation for Small Area Means under Non-normal Random Effects

Description: The empirical best linear unbiased predictor (EBLUP) is often used to produce small area estimates under the assumption of normality of the random effects. The exact mean squared error (MSE) for these approaches are unavailable, and thus must also be approximated. Staff will explore the use of estimating equations to obtain estimates of model parameters and the use of asymptotic expressions with a nonparametric bootstrap method to approximate the MSEs.

Highlights: During FY 2023, staff completed substantial revisions to a manuscript detailing the findings and conclusions from the work. Staff also completed additional simulation studies which investigated the T-distribution for the random effects and developed an appendix summarizing these additional findings. Staff further investigated the assumptions and asymptotic properties of estimating equations in the Fay-Herriot model.

Staff: Gauri Datta (x33426), Kyle Irimata, Jerry Maples, Eric Slud

B. Bayesian Hierarchical Spatial Models for Small Area Estimation

Description: Model-based methods play a key role to produce reliable estimates of small area means. Two popular models, namely, the Fay-Herriot model and the nested error regression model, consider independent random effects for the model error. Often population means of geographically contiguous small areas display a spatial pattern, especially in the absence of good covariates. In such circumstances spatial models that capture the dependence of the random effects are more effective in prediction of small area means. Staff members and external collaborators are currently developing a hierarchical Bayesian method for unit-level small area estimation setup. This generalizes previous work by allowing the area-level random effects to be spatially correlated and allowing unequal selection probability of the units in the sample.

Highlights: During FY 2023, staff collaborated with Hee Cheol Chung to explore effectiveness of a class of spatial random effects models as alternatives to the standard nested error regression model for unit-level small area estimation applications. Inspired by a joint research by a staff member with collaborator Chung that showed superiority of a class of spatial random effects models over the traditional Fay-Herriot model relying on independent random effects, staff proceeded to investigate if a similar superiority of spatial random effects models will continue to hold. Staff identified an application and prepared data that would be consistent with the setup of the models under consideration. Staff

conducted extensive simulations based on the setup of the above application. Preliminary simulations appear to confirm the effectiveness of some spatial random effects model. For both projects, the staff followed the Bayesian approach.

Staff: Gauri Datta (x33426), Ryan Janicki, Jerry Maples, Hee Cheol Chung (UNC Charlotte)

C. Exploration of Small Area Estimation via Compromise Regression Weights

Description: The empirical best linear unbiased predictor (EBLUP) is often used to produce small area estimates under the assumption of normality of the random effects. Model-based estimate of a small area mean is obtained by shrinking a "noisy" direct estimate to a regression synthetic estimate based on a model. If a model is mis-specified, model-based estimates of areas with less reliable direct estimates may be sub-optimal due to their overreliance on a poorly estimated model. Jiang et al. (2011, *JASA*) and Nicholas et al. (2020) proposed frequentist estimation of the model by minimizing an estimated total mean squared error (ETMSE). The method proposed by Jiang et al. is known in the literature as the "Observed Best Prediction" (OBP) method.

Highlights: During FY 2023, staff with his two former Ph.D. students Juhyung Lee and Jiacheng Li, developed a noninformative pseudo-Bayesian (later renamed as quasi-Bayesian) method based on the objective function key to the development of the OBP method. They have investigated robustness of their procedure, and the procedures based on OBP, and the standard EBLUPs. These three researchers jointly published a paper in *Journal of Survey Statistics and Methodology* based on their findings on this topic. Among the key findings, they found that their quasi-Bayesian method is (1) as good as the OBP method, (2) better equipped than the OBP method to estimate uncertainty of point estimates, and (3) as robust as the OBP method to mean misspecification in a simple example. But in realistic applications, none of these methods displayed robustness to omitted variables in regression. The EBLUPs are not robust to model misspecification even in the simple example for which the OBP and the quasi-Bayesian method enjoyed robustness. Staff continued development on the quasi-Bayesian method for spatial random effects model with Li. This work has been accepted a few months ago for a special issue of *Calcutta Statistical Association Bulletin*. In a sense, this research is a synthesis of one of the papers on Bayesian approach to spatial random effects model and another paper on quasi-Bayesian alternative to the OBP method based on Fay-Herriot model by the staff. Evaluation of the proposed method based on an application to estimation of four-person family median incomes for the contiguous U.S. states showed that the method is useful. Related simulation studies for this application reinforced this finding.

Staff: Gauri Datta (x33426), Juhyung Lee (University of Florida), Jiacheng Li (Wells Fargo)

D. Construction of Joint Credible Set of Ranks of Small Area Means

Description: This is a topic of great interest to the Census Bureau and many federal statistical agencies around the world. This project develops joint credible set of ranks of small area means based on an approximate highest posterior density credible set of small area means. This project creates joint posterior distribution of ranks of all small areas under consideration. The project also compares the performance of the Bayesian solution with the available frequentist solution. Staff is collaborating on this project with two external collaborators.

Highlights: During FY 2023, the staff devoted a considerable effort on construction of joint credible set of ranks of small area means. This is a topic of great interest to the Census Bureau and many federal agencies. This project develops joint credible set of ranks of small area means based on an approximate highest posterior density credible set of small area means. The project compares the performance of the Bayesian solution with the available frequentist solution. Staff is collaborating on this project with two external collaborators. Currently, they are using the proposed method for several applications, conducting simulations to compare performance of the Bayesian method with the frequentist method, and preparing a manuscript.

Staff: Gauri Datta (x33426), Abhyuday Mandal (University of Georgia), Yiren Hou (University of Michigan)

E. Machine Learning-assisted Small Area Estimation (SAE) Models

Description: The SAE models have been extensively employed for official federal statistics with a clear goal of predicting a response of interest in the sparse population. Most SAE models rely on a single parametric model that is subject to its predictive correctness. Because the nature of problem is prediction, our team has researched the applicability of machine learning methods, such as the ensemble of candidate prediction models and regression tree-based methods to enhance predictions for the SAE models. The staff has developed a tree-assisted Fay-Herriot (FH) model, which adopts the regression tree model to produce an optimized covariate set in order to improve the predictive performance of a conventional FH model.

Highlights: During FY 2023, staff researched and implemented two Bayesian versions of variable selection methods: 1) Stan programs with the Horseshoe parameter specification, and 2) C++ programs with the stochastic search variable selection procedure for small area

estimation to resolve the problem of high-dimensional covariates for the *Voting Rights Act* (Section 203) data sets.

Staff: Joseph Kang (x32467)

Spatial Analysis & Modeling

Motivation: It is often the case that data collected from large-scale surveys can be used to produce high quality estimates at large domains. However, data users are often interested in more granular domains or regions than can be reasonably supported by the data due to small samples which can lead to both imprecise estimates as well as unintended disclosure of respondent data. Indirect methods of inference which utilize statistical models, latent Gaussian processes, and auxiliary data sources have proven to be an effective method for improving the quality of published data products. In addition, there is often a high degree of clustering and spatial correlation present in these large data sets which can be exploited to improve precision. Statistical modeling can be used to incorporate spatial, multivariate, and temporal dependencies as well as to integrate various data sources to both improve quality as well as to produce new estimates in regions and sub-domains with sparse or no data.

Research Problems:
- Statistical methodology for integration of data from various sources.
- Development of unit-level models.
- Incorporation of survey weights in statistical models.
- Development of change-of-support methodology.
- Development of computationally efficient methods for fitting models to non-Gaussian data.
- Incorporation of spatially-correlated random effects in small area models.
- Model-based methods for prediction at low geographic levels.
- Mean-squared error, uncertainty, and interval estimation.
- Synthesis of privacy protection and model-based inference.
- Nonparametric covariance estimation.
- Inference for irregularly spaced observations from locally-stationary random fields.

Potential Applications:
- Estimation of health insurance coverage by different demographic classifications at different geographic levels.
- Creation of new custom tabulations of ACS data products.
- Improvement of the precision of noisy measurements of census counts or other variables subject to disclosure avoidance techniques.
- Methodology for producing public use synthetic micro data.

A. Change of Support Methodology
Description: We consider the problem of inference on a geographic region (target support) when observations correspond to one or more geographic regions (source support) which are distinct from the target support.

Highlights: During FY 2023, staff completed a paper describing methodology for producing estimates corresponding to a target support when observations on a distinct source support and are drawn from a differentially private noise distribution. It was shown how to incorporate auxiliary covariate. This paper was accepted for publication. Staff also applied this methodology to producing estimates on 'grids.' Staff considered the problem of estimation of the population 64+ in regularly-spaced 1 km^2 squares in Rhode Island.

Staff: Ryan Janicki (x35725), Andrew Raim, James Livsey, Kyle Irimata, Scott Holan (R&M)

B. Clustering Methods
Description: It is often the case that data are clustered by geographic region or by demographic characteristics. However, it may be difficult to ascertain the precise clustering mechanism or the variables on which clustering occurs. This project develops new methodology for automatic clustering of data for the purpose of producing accurate estimates with an emphasis on model-based methods using data collected under an informative survey design.

Highlights: During FY 2023, staff developed new unit-level models for sample survey data collected under an informative design. The likelihood (Gauss, Poisson, or Binomial) is exponentiated using the survey weights, and multivariate and spatial dependencies are introduced through a hierarchical Bayesian structure. A nonparametric Dirichlet process prior is then used to automatically cluster the data based on spatial and covariate information. The distributional theory was fully derived and code was developed. The code was compiled in preparation for submission as an R package, and a paper describing the methodology and computations for submission to a statistical software journal was started. Staff also began work of applying the methodology to a new ACS special tabulation.

Staff: Ryan Janicki (x35725), Paul Parker, Scott Holan (R&M)

C. Machine Learning and Spatial Modeling
Description: The use of auxiliary information such as covariate data and spatial structures in Bayesian hierarchical models is critically important for producing accurate predictions. However, it can often be the case that the quantity of available data is overwhelming, and

the number of potential predictors is far greater than the number of observations. In this setting it is challenging to select a manageable subset of predictors for use in a model, to specify a functional form for the relationship between the response and predictor variables, and to include all important interactions and correlations.

Highlights: During FY 2023, staff began a literature review of machine leaning and variable selection methodology. Staff collected American Community Survey, Demographic Household Characteristics (DHC), and decennial data and began experimentation using different methods for selecting a subset of the available data, the functional form of the selected subset, as well as all relevant interactions for the purpose of predicting different target tables, focusing on proposed 2020 S-DHC data products. Staff began work on creating a county-level S-DHC data set on which to test methodology.

Staff: Ryan Janicki (x35725), Kyle Irimata, James Livsey, Andrew Raim, Scott Holan (R&M)

D. Locally Stationary Spatial Processes
Description: Spatial processes observed over large domains often exhibit nonstationarity, which necessitates use of nonstationary covariance models. In this project, staff is developing a theoretical framework for locally stationary spatial processes and applying this methodology to related problems at the Census Bureau.

Highlights: During FY 2023, staff prepared an American Community Survey data set consisting of median household income and the number of housing units at the block group level to both illustrate the problem and to compare methodologies. Staff also began work on developing code for estimating the covariance function and the variogram and for predicting at locations where the process is not observed. Staff also considered application to the 'grids' project in which predictions need to be made on a 1km^2 lattice.

Staff: Soumen Lahiri, Ryan Janicki (x35725)

Time Series & Seasonal Adjustment

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

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

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

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

Highlights: During FY 2023, staff made progress on several projects, including: (a) writing of text and code for a book on multivariate real-time seasonal adjustment and forecasting, with new code and methodology for co-integration filter constraints, model. This is an e-book composed in Quarto; (b) further refined theoretical results and code for a mechanistic description of seasonality involving a marked point process. Developed several methods for modifying points, including random displacement, demonstrating the linkage between flow dynamics and point scattering; and (c) explored a new diagnostic for seasonality based upon forecast errors, with an assessment of both seasonal persistence and intra-seasonal association.

Staff: Tucker McElroy (240-695-3610; R&M), James Livsey, Osbert Pang, Anindya Roy

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

Highlights: During FY 2023, staff made progress on

several projects: (a) refined a method for comparing two specifications of differencing operator via multi-step ahead forecast mean squared error; (b) continued work on applying the quadratic forecast filter to recursively generate nonlinear processes, using ideas from Markov chain theory; (c) completed asymptotic results for polyspectral means, a type of estimator that involves a weighted integral of the polyspectral density. The exposition has been polished, and the paper submitted; (d) polished final results on maximal benefit of quadratic prediction over linear prediction; (e) completed a revision for a new procedure for local spectral density estimation that is optimal at the boundary of the frequency domain; (f) completed work on time series differential privacy, involving new notions of utility and privacy, and the use of all-pass filtering as a protective device. Extensions to multivariate time series are being developed; (g) formulated a measure of time series survey discontinuity, to assess whether a time series sample path has substantially changed after a change to survey methodology or data source; (h) completed a study of skip-sampling, a subsampling methodology formulated in the frequency domain; (i) completed draft of a paper on forecasting GDP growth via component trends and cycles; (j) continued numerical work to support methodology of instrumental variables in panel regression of forecast error on forecast revision; and (k) completed a project on the definition of locally stationary spatial processes, and the nonparametric estimation of their covariance structure.

Staff: Tucker McElroy (240-695-3610; R&M), James Livsey, Osbert Pang, Anindya Roy

Experimentation, Prediction, & Modeling

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

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

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

A. Developing Flexible Distributions and Statistical Modeling for Count Data Containing Dispersion
Description: Projects address myriad issues surrounding count data that do not conform to data equi-dispersion (i.e., where the (conditional) variance and mean equal). These projects utilize the Conway-Maxwell-Poisson (CMP) distribution and related distributions and are applicable to numerous Census Bureau interests that involve count variables.

Highlights: During FY 2023, staff wrote a technical report regarding advancements made to the COMPoissonReg package. Staff meanwhile researched the impact of the choice of prior distribution on the CMP-parametrized version of the COM-Poisson distribution through theoretical results and data simulations with varying sample sizes. Staff are now drafting a manuscript for peer review relating to the latter project. Finally, staff are studying various generalizations of the negative binomial distribution in order to compare and contrast their respective properties. To date, staff have developed R codes to study the generalized Conway-Maxwell Poisson (GCMP) distribution, including its probability function, cumulative distribution function, quantile function, and random number generator, along with codes to compute the mean, variance, and kth moment of the distribution.

Staff: Kimberly Sellers (x39808), Darcy Steeg Morris, Andrew Raim

B. Randomization, Re-randomization and Covariate Balance in Treatment-control Comparisons

Description: For comparing two treatments in a finite population setting, randomization is commonly employed in order to achieve covariate balance. The difference-in-means estimator is widely used for comparing the two treatments, and randomization based statistical inference can be carried out without making strong model assumptions. Both the estimator and the statistical inference can be improved by the appropriate use of covariates. Regression adjustment can in fact yield a more efficient estimator. Furthermore, possible covariate imbalance that could occur by chance can be mitigated by the use of re-randomization. Here the re-randomization is to be carried out repeatedly until covariate balance is achieved according to a specific criterion. These topics have received considerable attention in the recent and very recent literature.

Highlights: During FY 2023, staff reviewed the literature on regression adjustment, covariate balance and re-randomization. The available literature appears to be focused on simple random sampling and completely randomized experiments. Staff explored possible extensions to unequal probability sampling and the use of the Horvitz-Thompson estimator in place of the difference-in-means estimator, for treatment-control comparisons in embedded experiments. In particular, staff is exploring possible extensions to systematic samples, especially unequal probability systematic samples. This is motivated by some of the embedded experiments that were carried out by the Demographic Statistical Methods Division (DSMD) of the Census Bureau. Staff has been reviewing one of the embedded experiments carried out by the DSMD in order to explore potential applications of the proposed research. The experiment under review explored the effectiveness of a mailed prenotice letter on the response rates in the 2021 National Survey of College Graduates (NSCG) data collection. Staff also reached out to the DSMD staff in order to obtain the 2021 NSCG data. Staff propose to analyze the data in order to illustrate the application of the methodologies under development.

Staff also explored some of the variance estimation approaches currently being employed under the systematic sampling design. Because the actual variance cannot be estimated, variance estimation approaches such as successive difference replication (SDR) are based on using a sampling design that mimics the systematic sample. Since SDR is widely used, especially at DSMD, staff has been looking at properties of the corresponding variance estimate, and also exploring alternative variance estimation approaches in the context of equal probability systematic samples. Subsequently, staff plans to investigate variance estimation under the unequal probability systematic sampling design.

Staff: Emanuel Ben-David (x37275), Thomas Mathew

C. Bayesian Modeling of Privacy Protected Data with Direct Sampling

Description: This project investigates the direct sampler, first proposed by Walker et al. (*JCGS*, 2011), and its use in modeling data released via differential privacy. In particular, additive noise mechanisms based on Laplace, Double Geometric, and Discrete Gaussian distributions are considered. Here, inference must be carried out on noisy versions of statistics computed from sensitive data. The direct sampler may be used to draw the unobserved statistics as latent random variables within a Gibbs sampler, provided that conditionals take the form of weighted distributions which satisfy certain assumptions.

Highlights: During FY 2023, a manuscript on direct sampling with step functions was published in the journal *Statistics and Computing*. The manuscript features several applications - without differential privacy - which may be more familiar to a general audience of statisticians.

Staff: Andrew Raim (x37894)

D. Rejection Sampling for Weighted Densities

Description: This project investigates rejection sampling for weighted densities using proposals which relax the weight function. Weighted target distributions arise in many problems of interest, such as in posteriors or conditionals in Bayesian analysis which may not have a recognizable form. Here, exact sampling may be preferred to an MCMC method where draws are correlated, and it may be unclear whether chains have sufficiently mixed. A desirable proposal distribution is one which could be constructed (or adapted) to be arbitrarily close to the target - while maintaining a relatively low level of computational complexity - to yield a low probability of rejection.

Highlights: During FY 2023, staff presented initial results in several venues. Illustrations of the sampler explored univariate targets, including generation from the polynomial normal distribution and the posterior of the noise variance from a Gaussian Process. An initial implementation of the framework was completed in R. A manuscript on the initial work is in preparation. Staff explored additional applications of the approach, including rejection samplers for truncated multivariate normal distributions and weighted multivariate normal distributions (e.g., encountered as posteriors in Bayesian data analyses).

Staff: Andrew Raim (x37894), James Livsey, Kyle Irimata

Simulation, Data Science, & Visualization

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

Research Problems:
• Systematically develop an environment for simulating complex sample surveys that can be used as a test-bed for new data analysis methods.
• Develop flexible model-based estimation methods for sample survey data.
• Develop new methods for statistical disclosure control that simultaneously protect confidential data from disclosure while enabling valid inferences to be drawn on relevant population parameters.
• Investigate the bootstrap for analyzing data from complex sample surveys.
• Develop models for the analysis of measurement errors in Demographic sample surveys (e.g., Current Population Survey or the Survey of Income and Program Participation).
• Identify and develop statistical models (e.g., loglinear models, mixture models, and mixed-effects models) to characterize relationships between variables measured in censuses, sample surveys, and administrative records.
• Investigate noise multiplication for statistical disclosure control.

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

A. Visualizing the United States
Description: This project explores the structure and methods used to construct a visualization-based statistical atlas of the United States that reflects the life of Americans. Early statistical atlases produced by the Census Bureau from 1870 to 1920, as well as the more recent Census Atlas of the United States, provide inspiration for a modern format for both online and print. With a general audience in mind, the research investigates the design trade-offs between visualization for analysis and for presentation and the balance between maintaining statistical accuracy while engaging readers without professional statistical knowledge.

Highlights: During FY 2023, staff evaluated previous visualization work published by the Census Bureau, with an emphasis on communicating data to the public. Visual forms and formats were explored for a modern version of such work, especially for usage with higher granularity data. The purpose was to focus on individuals and make Census data more relatable to those who do not normally work with or read data. Graphics were published publicly, and staff noted the need to annotate charts and explain new visual forms to make charts useful to a wider audience.

Staff: Nathan Yau (CSRM, FLOWINGDATA.COM)

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

Highlights: During FY 2023, staff conducted research on statistical methods for protecting data confidentiality and respondent's privacy. Staff investigated statistical

properties of some recently proposed procedures for protecting data confidentiality. We assessed both disclosure risk and impacts on the accuracy of statistical inferences. Staff expects to write a paper reporting the findings of this study.

Staff: Tapan Nayak (x35191)

C. Frequentist and Bayesian Analysis of Multiply Imputed Synthetic Data

Description: Under this project, staff members will conduct research on some aspects of both frequentist and Bayesian analysis of multiply imputed synthetic data produced under a multiple linear regression model, synthetic data being created under two scenarios: posterior predictive sampling and plug-in sampling.

Highlights: During FY 2023, staff published the paper "Bayesian Analysis of Multiply Imputed Synthetic Data under Multiple Linear Regression Model" (Abhishek Guin, Anindya Roy, and Bimal Sinha) *International Journal of Statistical Sciences,* ISSN 1683-5603; Vol. 22(2), November 2022, pp 25-38 © 2022 Dept. of Statistics, University of Rajshahi, Bangladesh. This project is complete.

Staff: Bimal Sinha (x34890), Anindya Roy, Abhishek Guin (UMBC)

D. Bayesian Analysis of Singly Imputed Synthetic Data under a Multivariate Normal Model

Description: Under this project, staff members will conduct research on developing valid statistical inference about the mean vector and dispersion matrix under a multivariate normal model. The basic premise is that data are collected on a vector of continuous attributes all of which are sensitive and hence cannot be released and require protection. We assume synthetic data are produced under two familiar scenarios: plug-in sampling and posterior predictive sampling. In an earlier CSRM report, Klein and Sinha (2015) conducted frequentist analysis of the synthetic data. In this research Bayesian analysis of the synthetic data will be carried out.

Highlights: During FY 2023, a related paper has been accepted and it will appear in *International Journal of Statistical Sciences, Volume* 23, 2023. This completes the project.

Staff: Bimal Sinha (x34890), Anindya Roy, Abhishek Guin (UMBC)

E. Comparison of Local Powers of Some Exact Tests for a Common Normal Mean Vector with Unknown and Unequal Dispersion Matrices

Description: In this work, we consider the problem of constructing a confidence set for an unknown common multivariate normal mean vector based on data from several independent multivariate normal populations with unknown and unequal dispersion matrices. We provide a review of some existing exact procedures to construct a confidence set. These procedures can be readily used to construct exact tests for the common mean vector. A comparison of these test procedures is done based on their local powers. A large sample test procedure based on multivariate generalization of univariate Graybill-Deal estimate of the unknown mean vector is also considered. Applications include a simulated data set and also data from the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) 2021, conducted by the Bureau of the Census for the Bureau of Labor Statistics.

Highlights: During FY 2023, a paper has been submitted to the Probability/Statistics Section of the open access journal *Mathematics* (ISSN 2227-7390) for the Special Issue "Clustered Data Modeling and Statistical Meta-Analysis."

Staff: Bimal Sinha (x34890), Yehenew Kifle (UMBC), Alain Moluh (UMBC)

F. Analysis of Multiply Imputed Synthetic Data from a Univariate Normal Population

Description: This is a continuation of research reported in Klein & Sinha (2015). The problem is to draw valid inference about a univariate normal mean based on multiply imputed synthetic data generated under posterior predictive sampling. It turns out that the crux of the problem is to come up with suitable meta-analysis procedures in order to combine multiply imputed datasets. Various exact tests and one large sample test for the normal mean are discussed and a comparison is made based on their local powers. Two point estimates of the normal mean are also proposed and compared.

Highlights: During FY 2023, research was completed and a paper has been submitted for journal review.

Staff: Bimal Sinha (x34890), Biswajit Basak (University of Calcutta)

G. Analysis of One-Way ANOVA Model using Synthetic Data

Description: In this research, we consider the age-old ANOVA problem of testing the equality of means of several univariate normal populations with a common unknown variance, except that the data used for analysis arise from a synthetic version of the original observations. We address two versions of the synthetic data: one obtained under Plug-In sampling (PIS) method and the other under Posterior Predictive Sampling (PPS) method. We study its distributional properties (null and non-null) and provide enough computational details. A comparison of power is also provided. As expected, the power under the PIS method is more than that under the PPS method. A measure of privacy protection is also

evaluated and it turns out that the PIS method provides less protection than the PPS method, thus confirming the standard belief that accuracy of inference and privacy protection work in opposite directions!

Highlights: During FY 2023, research under the above topic was completed. The paper has appeared in the *CSRM Research Report Series.* We have also received referee's report from *Sankhya*, and the paper is being revised now for resubmission.

Staff: Bimal Sinha (x34890), Biswajit Basak (University of Calcutta)

H. Confidence Ellipsoids of a Multivariate Normal Mean Vector Based on Noise Perturbed and Synthetic Data with Applications

Description: In this research project we address the problem of constructing a confidence ellipsoid of a multivariate normal mean vector of a multinormal population based on a random sample from it. The central issue at hand is the sensitivity of the original data and hence the data cannot be directly used/analyzed. We consider a few perturbations of the original data, namely, noise addition and creation of synthetic data based on the plug-in sampling (PIS) method and the posterior predictive sampling (PPS) method. We review some theoretical results under PIS and PPS which are already available based on both frequentist and Bayesian analysis (Klein/Sinha, 2015, 2016; Guin/Roy/Sinha, 2023) and derive the necessary results under noise addition. A theoretical comparison of all the methods based on expected volumes of the confidence ellipsoids is provided. A measure of privacy protection (PP) is discussed and its formulas under PIS and PPS and noise addition are derived and the methods are compared. Applications to some Census Bureau data sets are illustrated.

Highlights: During FY 2023, research on this topic has been initiated and we hope to complete it within six months.

Staff: Bimal Sinha (x34890), Yehenew Kifle (UMBC), Biswajit Basak (Sister Nivedita University)

Summer at Census

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

Highlights: During FY 2023, staff organized the fourteenth annual *2023 SUMMER AT CENSUS* which brought 17 recognized scholars to the Census Bureau for 1-3 day (virtual) visits and cover broad topic themes including: community based-assessments, community resilience, measuring Asians in America, occupational restructuring, probability sampling, rural statistics, spatial & temporal statistics, statistical machine learning applications, survey methodology, survey sampling methodology, and the Understanding of America Study. Each scholar engaged in collaborative research with Census Bureau researchers and staff (Center for Statistical Research & Methodology; Associate Directorate for Demographic Programs; Social, Economic, Housing Statistics Division; Center for Behavioral Science Methods; Office of Strategic Alliance; Associate Directorate for Communications; Associate Directorate for Research & Methodology; Center for Economic Studies; Associate Directorate for Economic Programs; and Associate Directorate for Decennial Census Programs) on at least one current specific Census Bureau problem and presented a seminar based on his/her research.

Staff: Tommy Wright (x31702), Joseph Engmark

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: Joseph Engmark, Michael Hawkins, Kelly Taylor

3. PUBLICATIONS

3.1 JOURNAL ARTICLES (Peer-Reviewed), PUBLICATIONS

Aleshin-Guendel, S. and Steorts, R. (In Press). "Monitoring Convergence Diagnostics for Entity Resolution," *Annual Review of Statistics and Its Applications.*

Arsham, A., Bebu, I., and Mathew, T. (2023). "Cost-Effectiveness Analysis Under Multiple Effectiveness Outcomes: A Probabilistic Approach," *Statistics in Medicine, 42, 3936-3955*.

Basak J., Soliman A., Deo N., Haase, K., Mathur, A., Park, K., Steorts, R., Weinberg, D., Sahni. S., and Sanguthevar R. (2023). "On Computing the Jaro Similarity Between Two Strings," *Proceedings of the 19th International Symposium on Bioinformatics Research and Applications*, Springer, 31-44.

Datta, G.S., Lee, J., and Li, J. (In Press). "Pseudo-Bayesian Small Area Estimation," *Journal of Survey Statistics and Methodology.*

Datta, G.S. and Li, J. (In Press). "A Quasi-Bayesian Approach to Small Area Estimation Using Spatial Models," *Calcutta Statistical Association Bulletin.*

Deo, N., Sanguthevar R., Joyanta B., Soliman, A., Weinberg, D., and Steorts, R. (In Press). "Novel Blocking Techniques and Distance Metrics for Record Linkage," *Proceedings of The 25th International Conference on Information Integration and Web Intelligence (iiWAS), Lecture Notes in Computer Sciences*, Springer.

Guin, A., Roy, A., and Sinha, B. (2023). "Bayesian Analysis of Singly Imputed Synthetic Data under the Multivariate Normal Model," *International Journal of Statistical Sciences, Volume 23(2)*, November 2023.

Guin, A., Roy, A., and Sinha, B. (2022). "Bayesian Analysis of Multiply Imputed Synthetic Data under the Multiple Linear Regression Model," *International Journal of Statistical Sciences, Volume 22(2)*, 25-38.

Janicki, R., Holan, S. H., Irimata, K. M., Livsey, J., and Raim, A. (In Press). "Spatial Change of Support Models for Differentially Private Decennial Census Counts of Persons by Detailed Race and Ethnicity," *Journal of Statistical Theory and Practice.*

Kang, J., Morris, D.S., Joyce, P., and Dompreh, I. (In Press). "On Calibrated Inverse Probability Weighting and Generalized Boosting Propensity Score Models for Mean Estimation with Incomplete Survey Data," *Wiley Interdisciplinary Reviews (WIREs) Computational Statistics.*

Lucagbo, M. and Mathew, T. (2023). "Rectangular Tolerance Regions and Multivariate Normal Reference Regions in Laboratory Medicine," *Biometrical Journal, 65(3)*.

Lucagbo, M., Mathew, T., and Young, D. (2023). "Rectangular Multivariate Normal Prediction Regions for Setting Reference Regions in Laboratory Medicine*," Journal of Biopharmaceutical Statistics, 33(2), 191-209*.

Marchant, N.G., Rubinstein, B.I.P., and Steorts, R. (2023), "Bayesian Graphical Entity Resolution Using Exchangeable Random Partition Priors," *Journal of Survey Statistics and Methodology, 11*, 569-596.

McElroy, T. (2022). "Casting Vector Time Series: Algorithms for Forecasting, Imputation, and Signal Extraction," *Electronic Journal of Statistics, 16*, 5534-5569.

McElroy, T. (2022). "Stationary Parameterization of GARCH Processes," *Economics Bulletin, 42 (4).*

McElroy, T., Ghosh, D., and Lahiri, S. (2023). "Quadratic Prediction of Time Series via Autocumulants," *Sankhya A*, published online.

McElroy, T. and Jach, A. (2023). "Identification of the Differencing Operator of a Non-stationary Time Series via Testing for Zeroes in the Spectral Density," *Computational Statistics and Data Analysis*, 177, 107580.

McElroy, T. and Politis, D. (2022). "Optimal Linear Interpolation of Multiple Missing Values," *Statistical Inference for Stochastic Processes*, *25*, 471-483.

McElroy, T. and Politis, D. (2023). "Estimating the Spectral Density at Frequencies Near Zero," *Journal of the American Statistical Association*, published online.

McElroy, T., Roy, A., and Hore, G. (2023). "FLIP: a Utility Preserving Privacy Mechanism for Time Series," *Journal of Machine Learning Research, 24*, 1-29.

McElroy, T. and Trimbur, T. (2022). "Variable Targeting and Reduction in Large Vector Autoregressions with Applications to Workforce Indicators," *Journal of Applied Statistics*, *50*, 1515-1537.

Morris, D.S. and Raim, A.M. (2023). "Comparing Trial and Variable Association in Contingency Table Data Using Multinomial Models for Clustered Data." In *Proceedings of the 37th International Workshop on Statistical Modelling*. Dortmund, Germany: Statistical Modelling Society, 536-542.

Parker, P., Holan, S.H., and Janicki, R. (In Press). "Conjugate Modeling Approaches for Small Area Estimation with Heteroscedastic Structure," *Journal of Survey Statistics and Methodology.*

Parker, P., Holan, S.H., and Janicki, R. (2023). "A Comprehensive Overview of Unit Level Modeling of Survey Data for Small Area Estimation Under Informative Sampling," *Journal of Survey Statistics and Methodology, Vol 11, No. 4, 829-857.*

Parker, P., Holan, S.H., and Janicki, R. (2023). "Comparison of Unit Level Small Area Estimation Modeling Approaches for Survey Data Under Informative Sampling," *Journal of Survey Statistics and Methodology, Vol 11, No. 4, 858-872.*

Raim, A.M. (2023). "Direct Sampling with a Step Function," *Statistics and Computing, 33(22)*. https://doi.org/10.1007/s11222-022-10188.

Raim, A.M., Mathew, T., Sellers, K. F., Ellis, R., and Meyers, M. (2023). "Design and Sample Size Determination for Experiments on Nonresponse Follow-up using a Sequential Regression Model," *Journal of Official Statistics, 39(2),* 173-202.

Raim, A.M., Nichols, E., and Mathew, T. (2023). "A Statistical Comparison of Call Volume Uniformity Due to Mailing Strategy," *Journal of Official Statistics, 39,* 103-121.

Slud, E., Hall, A., and Franco, C. (In Press). "Small Area Estimates for Voting Rights Act Section 203(b) Coverage Determinations," *Calcutta Statistical Association Bulletin*.

Steorts, R. (2023). "A Primer on the Data Cleaning Pipeline," *Journal of Survey Statistics and Methodology, 11*, 553-568.

Wang, Z., Ben-David, E., and Slawski, M. (2023). "Regularization for Shuffled Data Problems via Exponential Family Priors on the Permutation Group." (Proceedings of the 26th International Conference on Artificial Intelligence and Statistics), *Proceedings of Machine Learning Research, Volume 206,* pgs 2939-2959. https://proceedings.mlr.press/v206/wang23a.

3.2 BOOKS/BOOK CHAPTERS

Mulry, M.H. and Mule, V.T. (2022). "Advances in the Use of Capture-Recapture Methodology in the Estimation of U.S. Census Coverage Error," In *Recent Advances on Sampling Methods and Educational Statistics. In Honor of S. Lynne Stokes.* Editors Hon Keung Tony Ng and Daniel F. Heitjan, 93–116, ISSN 2524-7735, https://doi.org/10.1007/978-3-031-14525-4

3.3 PROCEEDINGS PAPERS

Joint Statistical Meetings, American Statistical Association, Toronto, Ontario (Canada), August 5-10, 2023
2023 Proceedings of the American Statistical Association

- Virgile, M., Tuttle, A.D., Mathew, T., and Wang, L., "Exploring the Association between Training and Field Performance in 2020 U.S. Census Address Canvassing"
- Weinberg, D. and Thibaudeau, Y., "Orthogonal Designs for Computing the MLE of Discrete Latent Class Models in Record Linkage"

3.4 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY RESEARCH REPORTS

https://www.census.gov/topics/research/stat-research/publications/working-papers/rrs.html

RR (Computing #2022-01): Andrew Raim and Kimberly F. Sellers, "COMPoissonReg: Usage, the Normalizing Constant, and Other Computational Details," November 09, 2022.

RR (Statistics #2022-05): Kyle M. Irimata, Andrew M. Raim, Ryan Janicki, James A. Livsey, and Scott H. Holan, "Evaluation of Bayesian Hierarchical Models of Differentially Private Data Based on an Approximate Data Model," September 30, 2022.

RR (Statistics #2022-06): Eric Slud, Carolina Franco, Adam Hall, and Joseph Kang, "Statistical Methodology (2021) for Voting Rights Act, Section 203 Determinations," December 20, 2022.

RR (Statistics #2023-01): Mary H. Mulry, Cristina J. Tello-Trillo, Thomas Mule, and Andrew Keller, "Comparisons of Administrative Record Rosters to Census Self-Responses and NRFU Household Member Responses," February 21, 2023.

RR (Statistics #2023-02): Jose Asturias, William R. Bell, Rebecca Hutchinson, Tucker McElroy, and Katherine J. Thompson, "Building the Census Bureau Index of Economic Activity (IDEA)," March 10, 2023.

RR (Statistics #2023-03): Biswajit Basak and Bimal Sinha, "Analysis of One-Way ANOVA Model Using Synthetic Data," April 28, 2023 (revised October 17, 2023).

RR (Statistics #2023-04): Abhishek Guin, Anindya Roy, and Bimal Sinha, "Bayesian Analysis of Singly Imputed Synthetic Data under the Multivariate Normal Model," July 14, 2023.

RR (Statistics #2023-05): Tommy Wright, "Understanding Chao's Method of Probability Proportional to Size Sampling," August 17, 2023.

RR (Statistics #2023-06): Jiacheng Li and Gauri S. Datta, "A Quasi-Bayesian Approach to Small Area Estimation Using Spatial Models," September 6, 2023.

RR (Statistics #2023-07): Serge Aleshin-Guendel and Jon Wakefield, "Adaptive Gaussian Markov Random Fields for Child Mortality Estimation," September 7, 2023.

3.5 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY STUDY SERIES

https://www.census.gov/topics/research/stat-research/publications/working-papers/sss.html

SS (Statistics #2023-01): Andrew M. Raim and Elizabeth Nichols, "A Comparison of Map Usability via Bivariate Ordinal Analysis," April 12, 2023.

3.6 OTHER REPORTS

4. TALKS AND PRESENTATIONS

International Conference on Statistical Distributions and Applications (ICOSDA), Marshall University, Huntington, WV, October 13-15, 2022.

- Kimberly F. Sellers (Keynote Speaker), "Dispersed Methods for Handling Dispersed Count Data."

2022 Federal Committee on Statistical Methodology Research & Policy Conference, Walter E. Washington Convention Center, October 25-27, 2022.

- Joseph Kang, "Generating Survey Weights Using a Machine-learning Method and the Entropy Balancing Calibration Technique."

Kathryn Chaloner Lecture in the Statistical Sciences, Field of Dreams Conference, November 4, 2022.

- Kimberly F. Sellers (Featured Speaker), "You Count! (and, believe me, I should know!)."

Eighth International Conference on Statistics for Twenty-first Century - 2022, Kerala University, Kerala, India, December 16-19, 2022.

- Gauri S. Datta (Invited Talk, virtual), "Pseudo-Bayes Small Area Estimation."

16th International Conference on Computational and Financial Econometrics, King's College London, London, England, December 17-19, 2022.

- Tucker McElroy (virtual), "Multivariate Direct Filter Analysis for Co-integrated Processes."

Statistics Department Seminar, The George Washington University, February 10, 2023.

- Kimberly F. Sellers, "Regression Models for Count Data Containing Dispersion."

Statistical Sciences Department Seminar, North Carolina State University, March 27, 2023.

- Kimberly F. Sellers, "Dispersed Methods for Handling Dispersed Count Data."

Statistics and Biostatistics Departments Seminar, University of Washington, Seattle, WA, April 27, 2023.

- Kimberly F. Sellers, "Regression Models for Count Data Containing Dispersion."

Annual Data Science & Analytics Symposium- Sponsored by Bowie State University and U.S. Census Bureau, Bowie State University, Maryland, May 2-3, 2023.

- Tommy Wright, "Data at Foundation of America's Democracy."

2023 Data Science Leadership Summit, Boston University, Boston, Massachusetts, May 8-10, 2023.

- Tommy Wright, "Understanding Data and Their Uncertainty" (Panel on Developing an Effective Data Science Work Force).

2023 Symposium on Data Science and Statistics, St. Louis, Missouri, May 23-26, 2023.

- Andrew Raim, "Rejection Sampling for Weighted Densities by Majorization."

Eighth International Conference on Statistics for Twenty-first Century – 2022 (Annual meeting of the Statistical Society of Canada – 2023), Carleton University, Ottawa, Canada, May 28-31, 2023.

- Gauri S. Datta (Invited Talk, virtual), "Pseudo-Bayesian Small Area Estimation."

Annual meeting of the International Indian Statistical Association - 2023, Colorado School of Mines, Golden, CO, June 1-4, 2023.

- Gauri S. Datta (Invited Talk, virtual), "Pseudo-Bayesian Small Area Estimation using Spatial Models."
- Emanuel Ben-David, (Presented Short Course), "Data Analysis after Record Linkage: Sources of Error, Consequences, and Possible Solutions."

Government Advances in Statistical Programming (GASP), Virtual Conference, June 14-15, 2023.

- Tucker McElroy (virtual), "Analysis of Crisis Effect via Maximum Entropy."

2023 Summer Program in Research and Learning/Summer Program Advancing Techniques in the Applied Learning of Statistics, Georgetown University, June 16, 2023.

- Tommy Wright, "Data at Foundation of America's Democracy."

Eighth International Webinar on Recent Trends in Statistical Theory and Applications – 2023 (WSTA 2023), Kerala University, Kerala, India, June 29-July 02, 2023.

- Gauri S. Datta (Invited Talk, virtual), "A Quasi-Bayesian Approach to Small Area Estimation using Spatial Models."

2023 World Statistics Congress, Ottawa, Ontario (Canada), July 16-20, 2023.

- Tucker McElroy, "Constructing an Economic Index from API Time Series Data Updated with Nowcasts."
- Eric Slud, "Design-based Calibration Following Modeled (or Assisted) Base-weights."

The 37th International Workshop on Statistical Modeling, Dortmund, Germany, July 17-21, 2023.

- Darcy Morris, "Comparing Trial and Variable Association in Contingency Table Data using Multinomial Models for Clustered Data."

2023 Joint Statistical Meetings, American Statistical Association, Toronto, Ontario (Canada), August 5-10, 2023.

- Emanuel Ben-David, "Predicting Census Survey Response Rates via Interpretable Nonparametric Additive Models."
- Kyle Irimata, "Comparison of Measurement Error Models with Differentially Private Covariates."
- Joseph Kang, "On Calibrated Inverse Probability Weighting via a Machine Learning Model for Incomplete Survey Data."
- James Livsey, "sigexUI: An Object Orientated Framework for Structural Component Models."
- Tucker McElroy, "Characterizing Seasonality Through Multi-Step Ahead Forecasting Information Sets."
- Darcy Morris, "Machine Learning for Model Based Weighting Methods with Complex Nonresponse in the American Community Survey."
- Andrew Raim, "Statistical Modeling of Mobile Questionnaire Assistance in a Census."
- Anindya Roy, "Reconciliation of Univariate and Multivariate Time Series."
- Kimberly Sellers, "Metrics for Diversity and Inclusivity: A Review."
- Rebecca C. Steorts, "A Review of Statistical Record Linkage." Special Session Honoring Bill Winkler (with assistance by Dan Weinberg)
- Yves Thibaudeau, "Orthogonal Projection of the Extended MLE on the Parameter Space of Partially Estimable Log-Linear Models."
- Daniel Weinberg, "Implementing Winkler's Algorithm Using General Linear Models and Orthogonal Decompositions."

Statistical Colloquium, University of Maryland Baltimore Campus (Virtual), September 15, 2023.

- Tucker McElroy (Invited Talk), "A Framework for Time Series Aggregation and Seasonality Using Marked Point Processes."

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

Shane Lubold, University of Washington, "Consistently Estimating Network Statistics using Aggregated Relational Data," April 18, 2023.

Serge Aleshin-Guendel, Duke University, "Monitoring Convergence Diagnostics for Entity Resolution," April 19, 2023.

Ian Taylor, Colorado State University, "Fast Bayesian Record Linkage for Streaming Data Contexts," April 20, 2023.

Kimberly Sellers, Georgetown University/U.S. Bureau of the Census, "Dispersed Methods for Handling Dispersed Count Data," April 25, 2023.

Eric Slud & James Livsey, U.S. Bureau of the Census, "Approximating the Variances of Cell Estimates in a Multiway Contingency Table with Infused Noise and Partially Released Noisy Marginal Totals," May 2, 2023.

Jeremy Seeman (U.S. Census Bureau Dissertation Fellow), The Pennsylvania State University, "Flexible Formal Privacy for Public Data Curation," May 11, 2023.

Brady West, University of Michigan-Ann Arbor, *SUMMER (Virtually) AT CENSUS*, "Moving from a National Probability-Based Push-to-Web Survey to an Online Probability-Based Panel: What are the Issues?," May 30, 2023.

Xi Song, University of Pennsylvania, *SUMMER (Virtually) AT CENSUS*, "When Occupations Disappear: Inequality in Mobility Between Workers in Occupations with Job Expansion and Contraction," June 6, 2023.

Omar Perez Figueroa, University of California, Irvine, *SUMMER (Virtually) AT CENSUS*, "Understanding Resilience in Communities through Quantitative and Qualitative Approaches," June 8, 2023.

Andrew Raim (Joint work with James Livsey & Kyle Irimata), U.S. Bureau of the Census, "Rejection Sampling for Weighted Densities by Majorization," June 8, 2023.

Colm O'Muircheartaigh, Harris School of Public Policy/NORC at University of Chicago, *SUMMER (Virtually) AT CENSUS*, "Middle Alternatives, Acquiescence, and the Quality of Likert Scales," June 12, 2023.

Bijan Kimiagar, Citizen's Commitment Approach, *SUMMER (Virtually) AT CENSUS*, "Policy Research and Advocacy through Family Centered, Community-Based Assessments," June 14, 2023.

Brady West, University of Michigan-Ann Arbor, *SUMMER (Virtually) AT CENSUS,* "Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA), June 20, 2023.

Heather Kitada Smalley, Willamette University, *SUMMER (Virtually) AT CENSUS*, "Adjusting for Mode Effect in Surveys: The Challenges and Inferential Implications of Modeling Effect Type," June 22, 2023.

Douglas Nychka, Colorado School of Mines, *SUMMER (Virtually) AT CENSUS*, "Spatial Statistics Beyond the Textbook," July 11, 2023.

Soutir Bandyopadhyay, Colorado School of Mines, *SUMMER (Virtually) AT CENSUS*, "Regridding Uncertainty for Statistical Downscaling of Solar Radiation," July 12, 2023.

William Kleiber, University of Colorado Boulder, *SUMMER (Virtually) AT CENSUS*, "Two Problems in Statistical Climatology: High-Dimensional Processes and non-Gaussianity," July 13, 2023.

Cassandra Dorius, Iowa State University, *SUMMER (Virtually) AT CENSUS*, "Integrating People and Data to Promote Rural Vitality," July 17, 2023.

Bikas Sinha, Retired Professor of Statistics, Indian Statistical Institute, Kolkata, *SUMMER (Virtually) AT CENSUS*, "On Never-Ending Features of SRSWR (N, n) Designs and Related Unbiased Estimators for the Population Mean and Population Variance," July 20, 2023.

Sixia Chen, The University of Oklahoma, *SUMMER (Virtually) AT CENSUS,* "Analytic Tools for Handling Nonprobability Samples," July 26, 2023.

Greg Ridgeway, University of Pennsylvania, *SUMMER (Virtually) AT CENSUS*, "Scorecards, Benchmarking, and the Search for Unusual Hospitals, Communities, and Cops," July 28, 2023.

Marco Angrisani, University of Southern California, *SUMMER (Virtually) AT CENSUS*, "Recruitment, Retention, and Sample Representativeness in the Understanding America Study, a Probability-Based Online Panel," August 1, 2023.

Arie Kapteyn, University of Southern California, *SUMMER (Virtually) AT CENSUS*, "Collecting Person-Generated Health Data in a Population-Representative Panel," August 1, 2023.

Marco Angrisani, University of Southern California, *SUMMER (Virtually) AT CENSUS*, "Collection of Electronic Financial Records in a Population-Representative Panel," August 2, 2023.

Neil Ruiz, Pew Research Center, *SUMMER (Virtually) AT CENSUS*, "Solving the Gap on Asian Americans in Public Opinion Research," August 2, 2023.

Jerzy Wiecorek, Colby College, *SUMMER (Virtually) AT CENSUS*, "Design-Based Conformal Prediction for Survey Sampling," August 2, 2023.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

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

- **Yves Thibaudeau –** Record Linkage Analysis of Alternative Team "For … work defining 358 requirements for record linkage as used within the Census Bureau, identifying 44 of those requirements as the most critical, and evaluating 21 record linkage software packages with the critical requirements. (The Team) also evaluated four internal or open-source software packages for speed and quality of performance using demographic and business data sets…work advanced the Census Bureau's understanding of its record linkage capabilities and through (this) assessment generated candidate solutions to build a new platform."

- **Daniel Weinberg –** Record Linkage Analysis of Alternative Team "For … work defining 358 requirements for record linkage as used within the Census Bureau, identifying 44 of those requirements as the most critical, and evaluating 21 record linkage software packages with the critical requirements. (The Team) also evaluated four internal or open-source software packages for speed and quality of performance using demographic and business data sets…work advanced the Census Bureau's understanding of its record linkage capabilities and through (this) assessment generated candidate solutions to build a new platform."

Fellow, Association for Women in Mathematics

- **Kimberly Sellers –** For her work improving diversity and inclusion in the mathematical and statistical sciences through leadership positions in the American Statistical Association; for her leadership in the Joint Statistical Meetings, the Conference for Women in Statistics, and the Infinite Possibilities Conferences; and for her mentorship of early career women.

6.2 SIGNIFICANT SERVICE TO PROFESSION

Emanuel Ben-David

- Refereed papers for the 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023), *WIREs Computational Statistics, Mathematical Reviews, Journal of Statistics and Computing*
- Member, Committee for 2023 W. J. Dixon Award for Excellence in Statistical Consulting, American Statistical Association
- Member, Doctoral Defense Committee, Department of Statistics, George Mason University

Gauri Datta

- Associate Editor, *Sankhya*
- Associate Editor, *Journal of the Royal Statistical Society, Series A*
- Associate Editor, *Environmental and Ecological Statistics*
- Editorial Member, *Calcutta Statistical Association Bulletin*
- Refereed papers for *Journal of the Royal Statistical Society A, Sankhya, Survey Methodology, Journal of Survey Statistics and Methodology*
- Reviewer, Ph.D. in Statistics Dissertation, University of Rome, Italy

Kyle Irimata

- Refereed a paper for *Statistical Methods in Medical Research*

Ryan Janicki

- Refereed papers for *Journal of Statistical Software* and *Journal of Official Statistics*

Joseph Kang

- Associate Editor, *Biometrics and Biostatistics International Journal*
- Associate Editor, *Journal of Addiction and Prevention*
- Refereed papers for *Journal of the Royal Statistical Society: Series A* and *Statistical Methods in Medical Research*

Jerry Maples

- Refereed papers for *Journal of the Royal Statistical Society – Series A*

Thomas Mathew

- Associate Editor, *Sankhya*
- Associate Editor, *Journal of Multivariate Analysis*
- Associate Editor, *Journal of Occupational and Environmental Hygiene*
- Refereed papers for *Journal of Statistical Computation and Simulation, Journal of the Royal Statistical Society, Series C, Pharmaceutical Statistics, Statistical Methods in Medical Research, Statistics in Medicine*, and *BMC Medical Research Methodology*

Tucker McElroy

- Refereed papers for *Studies in Nonlinear Dynamics and Econometrics, Annals of Applied Probability, Survey Methodology, IEEE, Communications in Statistics, Journal of Time Series Analysis, Journal of the American Statistical Association, Econometrics and Statistics, Statistical Analysis and Data Mining, Statistica Neerlandica,, Journal of Nonparametric Statistics,* and *International Statistical Review*
- *Associate Editor, Journal of Time Series Analysis*
- *Guest Co-Editor, Journal of Official Statistics*
- Member, Scientific Committee, *9th Annual Conference of the International Association of Applied Econometrics*
- Member, Zellner Thesis Award Committee, Business & Economics Statistics Section, American Statistical Association
- Member, Council of Sections Representative, Business & Economics Statistics Section, American Statistical Association

Darcy Morris

- Associate Editor, *Communications in Statistics*
- Newsletter Editor, Survey Research Methods Section, American Statistical Association
- Program Chair-Elect, Government Statistics Section, American Statistical Association

Mary Mulry

- Associate Editor, *Journal of Official Statistics*
- Refereed a paper for *Journal of Official Statistics*

Tapan Nayak

- Associate Editor, *Journal of Statistical Theory and Practice*
- Guest Editor, Special Issue of *Statistics and Applications* in Memory of Professor C.R. Rao
- Refereed a paper for *Biometrika*

Andrew Raim

- Refereed papers for *Statistics and Computing*, *Communications in Statistics: Simulation and Computation*, *BMC Medical Research Methodology,* and *International Journal of Data Science and Analytics*

Kimberly Sellers

- Associate Editor, *The American Statistician*
- Commissioning Editor, *WIREs Computational Statistics*
- Refereed a paper for *Statistical Papers*
- Inaugural Chairperson (until 12/31/22) and Past Chairperson (beginning 1/1/23) Justice, Equity, Diversity, and Inclusion (JEDI) Outreach Group, American Statistical Association
- Vice-Chairperson, External Nominations and Awards Committee, American Statistical Association

Bimal Sinha

- Associate Editor, *Environmental Modeling and Assessment*
- Associate Editor*, Thailand Statistician*
- Associate Editor*, Calcutta Statistical Association Bulletin*
- Associate Editor, *Nepalese Journal of Statistics*

Eric Slud

- Associate Editor, *Journal of Survey Statistics and Methodology*
- Associate Editor, *Lifetime Data Analysis*
- Associate Editor, *Statistical Theory and Related Fields*
- Refereed papers for *Statistics in Transition - New Series, Annals of Statistics*, and *Communications in Statistics – Theory and Methods*

Rebecca Steorts

- Associate Editor, *Journal of Survey Statistics and Methodology*
- Associate Editor, *Journal of the American Statistical Association, Applications and Case Studies*
- Associate Editor, *Science Advances*
- Associate Editor*, Bayesian Analysis*
- Guest *Co*-Editor*, Journal of Survey Statistics and Methodology regarding special edition on data integration (with Joe Sakshaug)*

Yves Thibaudeau

- Refereed a paper for *Journal of Survey Statistics and Methodology*

Tommy Wright

- Refereed a paper for *Statistical Papers*
- Member*, The American Statistician* Editor Search Committee, American Statistical Association
- Member, Magic City Classic-Pioneers in STEM Discussion Panel: Reflections on the Past, a Glimpse of the Future in STEM (Oct 2022), Birmingham Civil Rights Institute

6.3 PERSONNEL NOTES

Serge Aleshin-Guendel (Ph.D. in biostatistics, University of Washington) accepted a research mathematical statistician appointment in our Spatial Analysis & Modeling Research Group.

Shane Lubold (Ph.D. in statistics, University of Washington) accepted a research mathematical statistician appointment in our Missing Data & Observational Data Modeling Research Group.

APPENDIX A **Center for Statistical Research and Methodology FY 2023**
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		
6550J06	Redistricting Data Program		
6550J08	Data Products Dissemination Preparation/Review/Approval		
6650J01	PES Planning and Project Management		
6650J20	2020 Evaluations-Planning and Project Management		
5350J01	Address Frame Updating Activities		
5350J04	Demographic Frame Updating Activities		
5450J06	Content, Forms Design, & Language		
5450J10	In-Person Enumeration Planning & Support		
5450J20	In-Office Enumeration Planning & Support		
5450J21	Response Data Quality		
5450J23	Response Processing Planning & Support		
5550J01	Data Products Creation & Dissemination		
5650J02	PES Planning & Project Management		
	1. Summary of Administrative Records Modeling for Some Enumerations in the 2020 Census	Mary Mulry	Tom Mule
	2. Comparisons of Administrative Records Rosters to Census Self-Responses and NRFU Household Member Responses	Mary Mulry	Tom Mule
	3. Supplementing and Supporting Non-Response with Administrative Records	Michael Ikeda	Tom Mule
	4. 2020 Census Privacy Variance	James Livsey	Phil Leclerc
	5. Experiment for Effectiveness of Bilingual Training	Andrew Raim	Renee Ellis
	6. Unit-Level Modeling of Master Address File Adds and Deletes	Eric Slud	Nancy Johnson
	7. Coverage Measurement Research	Jerry Maples	Tim Kennel
	8. Empirical Investigation of the Minimum Total Population of a Geographic District to Have Reliable Characteristics of Various Demographic Groups	Kyle Irimata	James Whitehorne
	9. Statistical Modeling to Augment 2020 Disclosure Avoidance System	Andrew Raim/Ryan Janicki	Michael Walsh
	10. Imputation Modeling for Multivariate Categorical Characteristic Data in 2030 Census	Darcy Morris	Tom Mule
	11. Mobile Questionnaire Assistance: Analysis and Simulation	Andrew Raim	Lisa Moore
	12. Exploring the Association between Training and Field Performance in 2020 U.S. Census Address Canvassing	Thomas Mathew	Lin Wang
	13. Assessments of the 2020 Nonresponse Follow-up Enumerator and Census Field Supervisor Training	Thomas Mathew	Lin Wang
	14. Agreements for Advancing Record Linkage	Rebecca Steorts	Krista Park
	15. Continuous Count Study	Michael Ikeda	Tom Mule
	16. Capture-Recapture Coverage Measurement using Administrative Records (Continuous Count Study)	Dan Weinberg	Tom Mule
	17. Cohort Component Birth Modeling (Continuous Count Study)	Emanuel Ben-David	Tom Mule
	18. Record Linkage Support for Decennial Census	Edward Porter	Aaron Gilary
6385J70	American Community Survey (ACS)		
	19. Voting Rights ACT (VRA) Section 203 Research Towards 2026 Determinations (also Decennial Project 6550J06)	Joseph Kang	James Whitehorne
	20. Assessing and Enhancing the ACS Experimental Weighting Approach Implemented in 2020 Data Products	Darcy Morris	Mark Asiala

TBA	**DEMOGRAPHIC** Demographic Statistical Methods Division (DSMD) Special Projects *21. Research on Biases in Successive Difference Replication Variance Estimation in Very Small Domains*.....	Eric Slud.....Tim Trudell
0906/1444X00	Demographic Surveys Division (DSD) Special Projects *22. Data Integration*.....	Edward Porter .....Christopher Boniface
7165023	Social, Economic, and Housing Statistics Division Small Area Estimation Projects *23. Research for Small Area Income and Poverty Estimates (SAIPE)* *24. Small Area Health Insurance Estimates (SAHIE)*	Jerry Maples.....Wes Basel Ryan Janicki.....Wes Basel
1183X01 1183X90	**ECONOMIC** General Economic Statistical Support General Economic Statistical Program Management *25. Use of Big Data for Retail Sales Estimates*..... *26. Seasonal Adjustment Support*..... *27. Seasonal Adjustment Software Development and Evaluation* ..... *28. Research on Seasonal Time Series - Modeling & Adjustment Issues*..... *29. Supporting Documentation & Software for Seasonal Adjustment* *30. Exploring New Seasonal Adjustment & Signal Extraction Methods*..... *31. Small Area Estimation for the Annual Integrated Economic Survey*.....	Darcy Morris.....Stephen Kaputa Tucker McElroy... Kathleen McDonald-Johnson James Livsey..... Kathleen McDonald-Johnson Tucker McElroy... Kathleen McDonald-Johnson James Livsey..... Kathleen McDonald-Johnson James Livsey.....Colt Viehdorfer Jerry Maples.....Jenny Thompson
0331000	**PROGRAM DIVISION OVERHEAD** *32. Research Computing*.....	Chad Russell ..... Jaya Damineni
9401021	**NATIONAL CANCER INSTITUTE** *33. Modeling Tobacco Use Outcomes with Data from Tobacco Use Supplement – Current Population Survey*.....	Isaac Dompreh ..... Benmei Liu

APPENDIX B

FY 2023 PROJECT PERFORMANCE MEASUREMENT QUESTIONNAIRE

CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY

Dear

As a sponsor for the FY 2023 Project described below, please (1) provide feedback on the associated Highlights Results/Products by responding to the questions to the right, (2) sign, and (3) return the form to Tommy Wright.

Your feedback will be shared with _____ to improve our future collaborative research.

_____
Tommy Wright/Chief, CSRM

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

Brief Description of Results/Products from FY 2023 ***(CSRM Contact will provide)****:*

TIMELINESS:
Established Major Deadlines/Schedules Met

1. Were all established major deadlines associated with this project or subproject met?

□ Yes □ No □ No Established Major Deadlines

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

2. Were there any improved methods, developed techniques, solutions, or new insights offered or applied on this project or subproject in FY 2023 where a CSRM staff member was a significant contributor?

□ Yes □ No

3. Are there any plans for implementation of any of the improved methods, developed techniques, solutions, or new insights offered or applied on this project?

□ Yes □ No

OVERALL:
Expectations Met

4. Overall, the CSRM efforts on this project during FY 2023 met expectations.

□ 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 Signature Date

Center for Statistical Research and Methodology

Research & Methodology Directorate

STATISTICAL COMPUTING AREA
Joseph Kang (Acting)

Record Linkage & Machine Learning Research Group
Yves Thibaudeau
- Emanuel Ben-David
- Xiaoyun Lu
- Rebecca Steorts
- Dan Weinberg

Missing Data & Observational Data Modeling Research Group
Darcy Morris
- Isaac Dompreh
- Shane Lubold
- Jun Shao (U. of WI)

Joseph Kang

Research Computing Systems & Applications Group
Chad Russell
- Tom Petkunas
- Ned Porter

Simulation, Data Science, & Visualization Research Group
Tommy Wright (Acting)
- Bimal Sinha (UMBC)
- Nathan Yau (FLOWINGDATA.COM)

MATHEMATICAL STATISTICS AREA
Eric Slud

Sampling Estimation & Survey Inference Research Group
Eric Slud (Acting)
- Mike Ikeda
- Patrick Joyce
- Mary Mulry
- Tapan Nayak (GWU)

Small Area Estimation Research Group
Jerry Maples
- Gauri Datta
- Kyle Irimata

Spatial Analysis & Modeling Research Group
Ryan Janicki
- Serge Aleshin-Guendel
- Soumendra Lahiri (Washington U.)
- Paul Parker (U. of CA, Santa Cruz)

Time Series & Seasonal Adjustment Research Group
James Livsey
- Osbert Pang

Tucker McElroy (Acting)
- Anindya Roy (UMBC)

Experimentation, Prediction, & Modeling Research Group
Tommy Wright (Acting)
- Thomas Mathew (UMBC)
- Andrew Raim
- Kimberly Sellers (NC State U.)

OFFICE OF THE CHIEF
Tommy Wright
- Kelly Taylor
- Joe Engmark
- Adam Hall
- Michael Hawkins