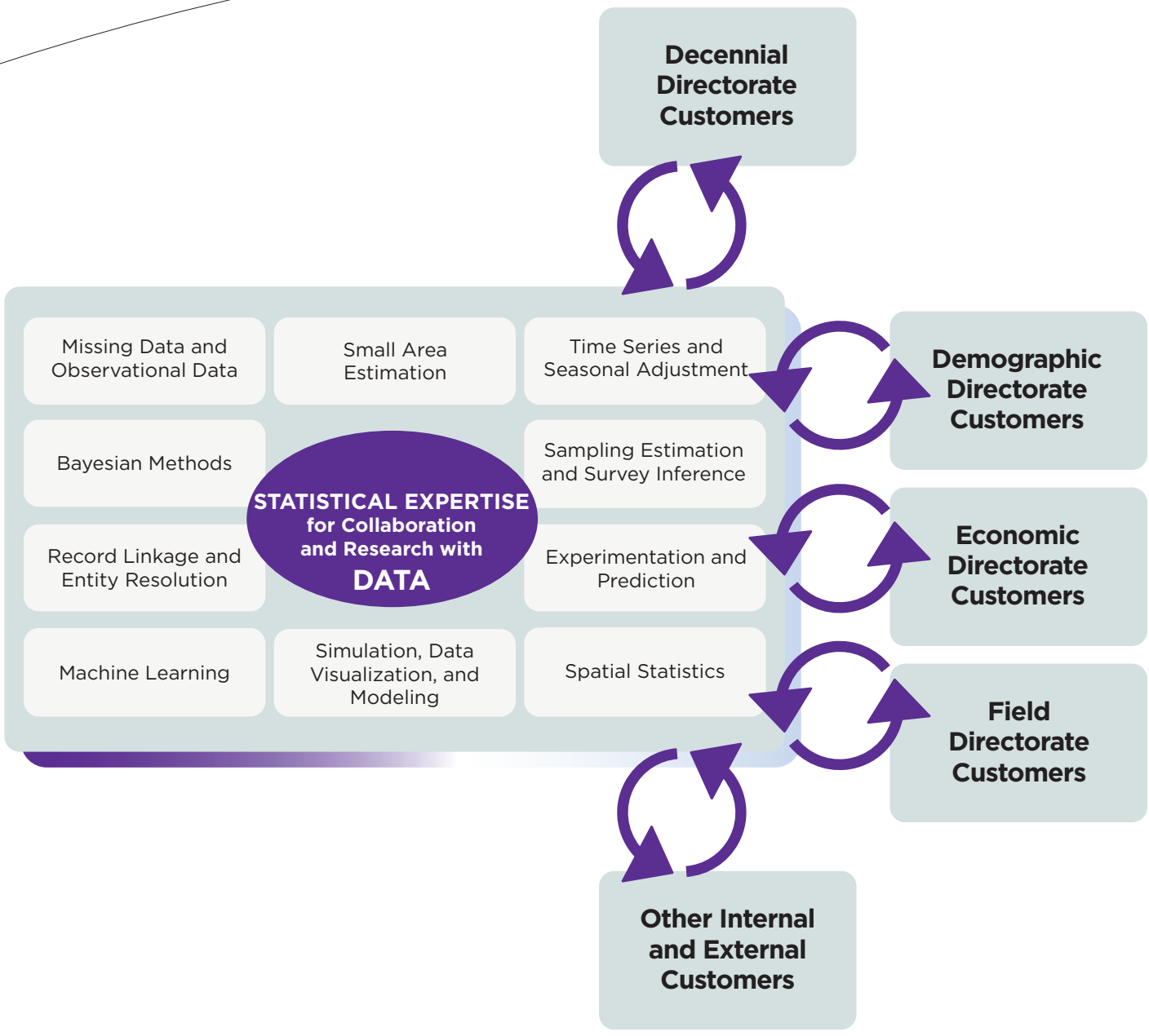
Annual Report of the Center for Statistical Research and Methodology

Research and Methodology Directorate

Fiscal Year 2022



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 2022, 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 2022 follow, and more details are provided within subsequent pages of this report:

- Published research results motivated by the Census Bureau's 2019-2020 Tracking Survey, conducted in two modes as a telephone probability sample survey and also as a nonprobability web sample survey. The report studies methodological issues concerning household sample survey estimation following weight-adjustment. New design-based theory provided in this report justifies generalized raking in settings where the correct weights (defined here for the first time in a design-based framework) satisfy a parametric model, and large-sample theory is established for adjusted-weight sample survey estimators and their variance estimates whether such a model holds or not. [CSRM (Slud, Morris)]

- Completed two manuscripts reporting statistical modeling to augment the 2020 Disclosure Avoidance System. The first manuscript considers the use of spatial change-of-support models to produce counts on small geographies. The second manuscript shows when a continuous Gaussian distribution provides a suitable approximation for the Laplace or Discrete Gaussian noise mechanism in a Bayesian hierarchical model. [CSRM (Raim, Janicki, Irimata, Livsey); R&M (Holan)]

- Produced a research paper (with colleagues in the Economic Statistical Methods Division, Economic Indicators Division, and Research & Methodology Directorate) describing a hierarchical Bayesian mass imputation model methodology for estimating state-level retail sales based on data from a third-party aggregation. An imputation model is built using the third-party data and applied to obtain imputations for all establishments in the sample 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 paper's aim 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. [CSRM (Morris), ESMD (Kaputa, Thompson); EID (Hutchinson)]

- Made research progress on several seasonal time series modeling and adjustment issues: (a) continued assessment of benchmarking optimization methodology for indirect adjustment of quarterly timeseries; (b) developed nowcasting methodology for an index of economic activity, based on Census time series; (c) applied benchmarking techniques to constituent variables of GDP, in order to remove residual seasonality while preserving aggregation relation; and (d) developed mathematical tree structure of aggregation relations and extended benchmarking code into a top-down algorithm. [R&M (McElroy, Bell); CSRM (Livsey, Pang)]

- Organized and presented a series of lectures aiming to provide deep knowledge in statistical methods for compensating for missing data – "Statistical Methods for Handling Incomplete Data." Twenty-three Census Bureau staff members participated in the four (90 minutes each) lectures: Center for Economic Studies (1); Center for Statistical Research & Methodology (6); Demographic Statistical Methods Division (4); Decennial Statistical Studies Division (6); Economic Statistical Methods Division (4); Research & Methodology Directorate (1); and Social, Economic, & Housing Statistics Division (1). [CSRM (Wright, Morris, Shao)]

- Published on the Census Bureau's website two "interactive" Research Data Visuals as part of The Ranking Project (Comparisons of A State with Each Other State / Estimated Rankings of All States). For each of 88+ Topics and the years 2018 and 2019 (now also 2021) based on published/official American Community Survey 1-year data, the first visual shows statistical comparisons of a state with each of the other states. For the same Topics, years, and data, the second visual shows statistical uncertainty in the overall estimated ranking of all fifty-one states (includes DC) using a novel joint confidence region (Klein, Wright, Wieczorek, 2020). [CSRM (Wright, Hall); FLOWINGDATA.COM/CSRM (Yau); Colby College (Wieczorek)]

- Received journal acceptance of manuscript that explores effectiveness of various spatial random effects models that capture dependence of the random effects as alternatives to the Fay-Herriot model. [CSRM (Datta, Janicki, Maples)]

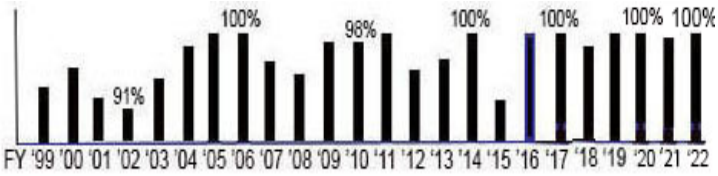
- Finalized for publication a book chapter [for volume in honor of S. Lynn Stokes retirement, former Census Bureau employee—Advances on Sampling Methodology and Educational Statistics (2023, Springer)] highlighting a discussion of the Post Enumeration Survey methodology and the evolution of its implementation to evaluate census coverage at the U.S. Census Bureau. [CSRM (Mulry); DSSD (Mule)]

How Did We[1] Do...

For the 24th year, we received feedback from our sponsors. Near the end of fiscal year 2022, our efforts on 37 of our programs (Decennial, Demographic, Economic, Administration, External) sponsored projects/subprojects with substantial activity and progress and sponsor feedback (Appendix A) were measured by use of a Project Performance Measurement Questionnaire (Appendix B). Responses to all 37 questionnaires were obtained with the following results (The graph associated with each measure shows the performance measure over the last 24 fiscal years):

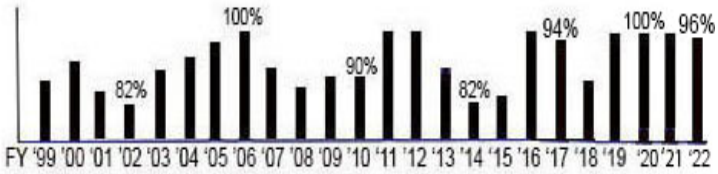
Measure 1. *Overall, Work Met Expectations*

Percent of FY2022 Program Sponsored Projects/Subprojects where sponsors reported that overall work met their expectations (agree or strongly agree) (37 out of 37 responses) ……. 100%



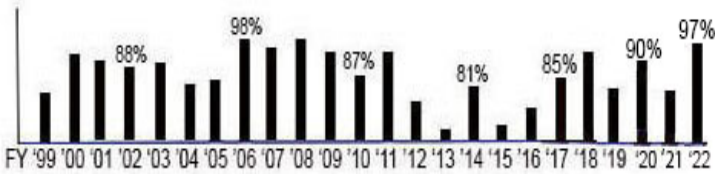
Measure 2. *Established Major Deadlines Met*

Percent of FY2022 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met (22 out of 23 responses) ……..…..…..…..…..96%



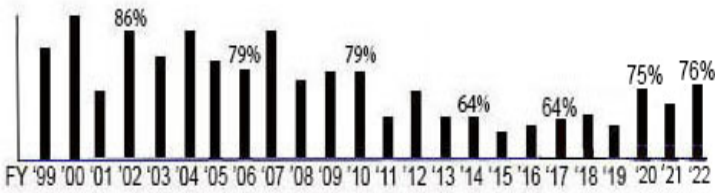
Measure 3a. *At Least One Improved Method, Developed Technique, Solution, or New Insight*

Percent of FY2022 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight (36 out of 37 responses) ……..… 97%



Measure 3b. *Plans for Implementation*

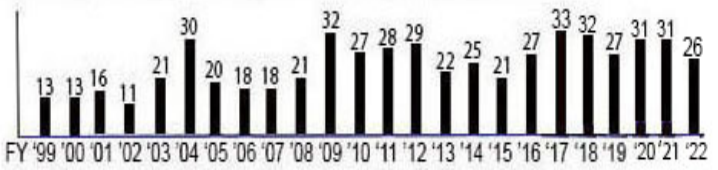
Of these FY2022 Program Sponsored Projects/Subprojects reporting at least one improved method, technique developed, solution, or new insight, the percent with plans for implementation (28 out of 37 responses) ……..…….. 76%



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

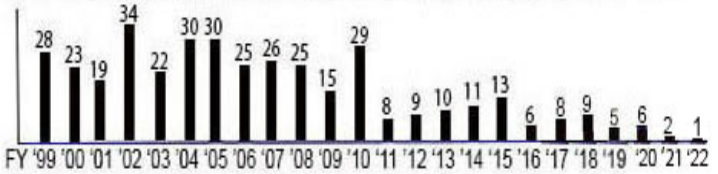
Measure 4. *Journal Articles, Publications*

Number of peer reviewed journal publications documenting research that appeared (18) or were accepted (8) in FY2022 …….. 26



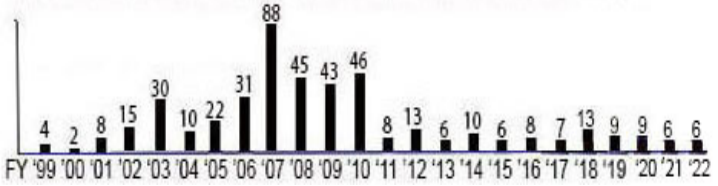
Measure 5. *Proceedings, Publications*

Number of proceedings publications documenting research that appeared in FY2022 ……..……. 1



Measure 6. *Center Research Reports/Studies, Publications*

Number of center research reports/studies publications documenting research that appeared in FY2022 ……..… 6



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

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

TABLE OF CONTENTS

1. COLLABORATION

1.1 DATA CODING/EDITING/IMPUTATION (Decennial Project 6550H01)

1.2 REDISTRICTING DATA PROGRAM (Decennial Project 6550H06)

1.3 DATA PRODUCTS DISSEMINATION PREPARATION/REVIEW/APPROVAL (Decennial Project 6550H08)

1.4 PES PLANNING & PROJECT MANAGEMENT (Decennial Project 6650H01)

1.5 2020 EVALUATIONS – PLANNING & PROJECT MANAGEMENT (Decennial Project 6650H20)

1.6 ADDRESS FRAME UPDATING ACTIVITIES (Decennial Project 5350H01)

1.7 DEMOGRAPHIC FRAME UPDATING ACTIVITIES (Decennial Project 5350H04)

1.8 CONTENT, FORMS DESIGN, & LANGUAGE (Decennial Project 5450H06)

1.9 IN-PERSON ENUMERATION PLANNING & SUPPORT (Decennial Project 5450H10)

1.10 IN-OFFICE ENUMERATION PLANNING & SUPPORT (Decennial Project 5450H20)

1.11 RESPONSE DATA QUALITY (Decennial Project 5450H21)

1.12 RESPONSE PROCESSING PLANNING & SUPPORT (Decennial Project 5450H23)

1.13 PES PLANNING & PROJECT MANAGEMENT (Decennial Project 5650H02)

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 FY2022, staff in our center collaborated with staff in Decennial Statistical Studies Division to prepare a paper about the use of administrative records to enumerate some households in the 2020 Census. After internal review, a draft of the paper will be submitted to a journal.

The paper shows how the Census Bureau incorporated the use of administrative records and other innovations into the 2020 Census. The innovation of using administrative records to enable classifying some addresses as occupied, vacant, or nonresidential when neither a self-response nor NRFU response was available. For occupied housing units where possible, the processing created a roster of the residents with their available characteristics that could be used for enumeration.

The COVID-19 pandemic caused the Census Bureau to delay NRFU interviewing originally planned for May 13 to July 24 of 2020 to be conducted from July 16 to October 15. In addition, the pandemic caused IRS to delay its deadlines for companies to send W2 and 1099 forms to taxpayers and to delay the deadline for filing a tax return from April 15 to July 15. Fortunately, the research concerning the use of ARs for census enumeration started in 2012 and enabled the Census Bureau to provide alternative paths to enumerate some households. The Census Bureau's AR modeling staff and subject matter experts were able to create adaptations to the planned procedures for AR enumeration to allow extended use of ARs for enumeration in ways that had not been planned. The innovation of using ARs to enumerate some households created the capability and

the flexibility to implement AR processing, and this was an essential factor in the success of the 2020 Census.

Staff: Mary Mulry (682-305-8809)

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

Description: The Census Bureau 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 FY2022, staff in our center has continued to collaborate with staff in the Decennial Statistical Studies Division (DSSD) and Center for Economic Studies (CES) on a study that compares 2020 Census rosters from self-responses and NRFU household member responses to AR rosters at addresses where both are available, and they differ on the household size. The authors intend to submit a version of the paper to a journal.

The evaluation project has focused on providing information about AR modeling and AR enumeration that will aid in planning the 2030 Census. In particular, the research concentrated on addresses that had both an AR roster and a census roster and examined whether the two rosters agreed on household size. The analysis was restricted to census self-responses and Nonresponse Follow-up (NRFU) household member responses because these are the highest quality census responses. The census rosters were partitioned by the timing of the submission of the response, which were before July 30 and after July 30 for self-responses and after July 30 for NRFU household member responses. Also considered was the quality category of the AR rosters available for enumerating the address, which are One-Visit Multiple Source, Closeout Multiple Source, and Closeout Household Size Only, listed in the order of their quality.

The combined results of the study indicate 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 AR roster. In addition, the results show the value ARs can bring to improvements in census enumeration methodology. This information has the potential of being useful in the search to find better ways to account for these types of differences as part of the 2030 Census planning and may contribute to identifying other sources and rostering methods to use in the future.

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 FY2022, staff in our center collaborated with staff in the Decennial Statistical Studies Division to prepare a paper in response to an invitation to submit a paper to a volume being published in honor of Dr. S. Lynne Stokes on her retirement from Southern Methodist University in the summer of 2022. The title of the volume is *Emerging Topics in Statistics and Biostatistics*. Dr. Stokes is a former Census Bureau employee who has contributed to the methodology for data collection and estimation for the Post Enumeration Survey (PES) at different points in her career. The discussion of the PES methodology and the evolution of its implementation to evaluate census coverage at the U.S. Census Bureau includes descriptions of her contributions to the design of quality control of the interviewing for the interviewing for the PES. Also included is a discussion of Dr. Stokes' development of an estimator of the residual fabrication in the final data from the PES interviews. These contributions were important information used in the decision about whether to use the data collected in the 1990 PES to construct an adjustment of the 1990 Census for coverage error. The decision by the Census Bureau was not to adjust the 1990 Census, and that decision was challenged in court but ultimately

was upheld by the U.S. Supreme Court on the basis that the U.S. Constitution requires the Census to be an 'actual enumeration.'

The title of the paper is "Advances in the Use of Capture-Recapture Methodology in the Estimation of U.S. Census Coverage Error." 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 methods 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. Results from the 2020 PES are not included because they were not expected to be available in time. The paper also includes a short discussion about research on the use of administrative records census enumeration. The volume is expected to be published in November 2022.

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 FY2022, seven documents written by staff were released as *Decennial Statistical Studies Division 2020 Decennial Census Memoranda A-16 and A-17. Memorandum A-16* "2020 Census Data Quality Administrative Record Usage Analysis: Initial Evaluation of the Puerto Rico Tax File for AR Modeling" is as described in the title. *Memorandum A-17* "2020 Census Data Quality Administrative Record Usage Analysis: Application of an Unsupervised Outlier Detection Methodology to the 2020 Decennial Census Master Address File" includes six attachments, Attachment 1 looks at methodology and basic results, Attachment 2 compares tract-level results to 2018 IRS 1040 response rates and differences between 2019 and 2018 IRS 1040 response rates, Attachment 3 does an overall comparison of results to the September production AR modeling results, Attachment 4 does a comparison focused on results for AR modeling deletes, Attachment 5 does a comparison focused on AR modeling deletes by ratio object score category (categories based on quantiles of the score from the outlier detection methodology), and Attachment 6 looks at AR modeling deletes and vacants by whether they were former vacant/delete overlap cases.

Staff also wrote four additional draft memoranda related to the outlier detection methodology. The first memorandum compares outlier detection results for AR modeling closeout deletes and closeout vacants to AR modeling deletes and vacants. AR modeling vacants and closeout vacants tend to have similar distributions of ratio object scores and outlier detection modeling variables while AR modeling closeout deletes tend to have lower ratio object scores and distributions of individual modeling variables shifted away from values suggesting questionable unit status when compared to AR modeling deletes. The second compares outlier detection results to AR modeling results for Puerto Rico. Because the AR modeling did not assign unit status in Puerto Rico, a pseudo-status was assigned using the occupied, vacant, and delete removal flags. Units with a pseudo-status of delete tended to have the highest ratio object scores and individual outlier detection modeling variables shifted toward values suggesting questionable unit status. The reverse was true for units with a pseudo-status of occupied. The third compares outlier detection results for AR modeling deletes that were assigned by AR on the Census Unedited File (CUF) to those for AR modeling deletes not assigned by AR on the CUF. The fourth does a similar comparison for AR modeling closeout deletes assigned by AR and not assigned by AR. In both cases the outlier detection results are similar for units assigned by AR and units not assigned by AR.

Staff also fit five multinomial logistic regression models on 2010 AR HU data for non-response follow-up (NRFU) units (with 2010 Census Unedited File (CUF) housing unit (HU) size as the dependent variable) and applied the models to the 2020 AR modeling HU data. The model results were then compared to the 2020 CUF results (the analysis generally excludes Puerto Rico). The predicted HU size was the HU size with the highest predicted probability from the model. Units with the highest predicted probability (for their predicted HU size) among "eligible" units (either NRFU response or with HU status assigned by AR modeling) were "selected" and the CUF HU size of "selected" units was replaced with the predicted HU size for the purpose of calculating population totals. Two different numbers of "selected" units were used for each model (results for the smaller number had already been obtained for four of the models). Staff documented the results from the five models in two draft memoranda and five new outlines, one for each of the models. The new outlines for the four models which had been run previously show results for the "selection" of the larger number of units. The outline for the fifth model shows results for both "selections."

The first draft memorandum shows results for the "initial" model with separate sub-models for different values of the relevant IRS 1040 household size, the second draft memorandum shows results for the model based on AR household composition. The other three models include a

"combined" model based on IRS 1040 household size but without separate sub-models for different values, a model based on AR household size, and a new model that is basically the same as the "initial" model when current year IRS 1040 household size is positive but is based on AR household size when AR household size is positive but current year IRS 1040 household size is zero. The "initial" model mostly does somewhat better both in agreement to the CUF population total for selected units (both all selected units and selected units where status is not assigned by AR modeling) and in accurate prediction of the CUF HU size (for selected units where HU status was not assigned by AR modeling). The results from the "initial" model are similar to those from the fifth model.

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 FY2022, multiple simulation configurations were set up, run, and analyzed. This includes setup of bottom-up and top-down workload sequences. These setups provided insight to reduction of variance across successive noise replications. Our initial finding is that bottom-up sequences are more effective for our purpose than top-down. These experiments lead to the exploration of the ordering of margins (rows in a design matrix) when performing L1-regression. The conjecture that distance between DP marginal sums and released DP margins effect L1-regression results was explored through a timing study.

The team investigated the computational speed gains from initializing the L1 regression problem in advantageous regions of the parameter manifold. This required modifying existing library functions to allow inclusion of initial coefficient estimates. Additionally, the number of slack vs non-slack constraints were reexamined for successive sequences of bottom-up workloads. This entailed accounting which marginal and details constrains switched from slack to non-slack (and vice versa).

Staff: James Livsey (x33517), Eric Slud

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

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

Highlights: During FY2022, staff worked with Decennial Statistical Studies Division and Center for Economic Studies colleagues to prepare final reports on adaptation of models for identifying and enumerating occupied housing units on American Indian Reservations; and adaptation of models for identifying and enumerating occupied housing units for off-campus college/university housing units. This project has concluded; ideas form the foundation for related 2030 Decennial Census research.

Staff: Darcy Steeg Morris (x33989), Yves Thibaudeau

E. 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 FY2022, staff explored records from the 2020 Census nonresponse follow-up operation on the Enterprise Data Lake and carried out initial analyses using regression based on the continuation-ratio logit model. Some aspects of the original experimental protocol were discovered to not have been followed in the field; staff considered adjustments to the originally proposed analysis to account for deviations.

Staff: Andrew Raim (x37894), Thomas Mathew, Kimberly Sellers, Renee Ellis (CBSM), Mikelyn Meyers (CBSM), Luke Larson (CBSM)

F. 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 FY2022, there was no progress. Activity resumes in FY2023 to create a brief Final Report.

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

G. Record-Linkage Support for the Decennial Census

Description: The Census Bureau is exploring avenues to support or replace traditional enumeration processes for the population decennial census by vastly expanding the use of administrative records, and publicly or commercially available data sources. A decennial project is tasked with researching all the aspects as well as the full potential of using data lists to improve data quality, guarantee confidentiality and cut costs. In particular, this entails a thorough research of record-linkage methods and software packages as well as of the many datasets available from governmental, public and commercial sources. A comprehensive "reference file" or "reference database" is under construction which will include individuals found in multiple administrative records sources other than the Numident File from the Social Security Administration or file(s) from the Internal Revenue Service. The timing of the project is ahead of the traditional decennial research cycle and concrete options for the 2030 Census are anticipated.

Highlights: During FY2022, staff explored and rated commercial and internal/open source offerings for record linkage in the Analysis of Alternatives (AoA) assessment. Internal and open source solutions were more finely tested on Census data in the Quantitative and Qualitative Investigations (QaQIs). After careful consideration, the solutions tested were *BigMatch*, *SASMatch*, *FastLink*, and MAMBA. Three matching exercises were attempted: matching the 2010 Census Unedited File (CUF) to the 2010 Census Coverage Measurement (CCM) P-sample, deduplicating the 2010 CUF, and matching the Business Register (BR) to the Small Business Administration (SBA) Disaster Loan Data. All solutions were rated qualitatively as well as by quantitatively on metrics such as speed, accuracy, precision, recall, and F1-score. At the end of the testing process, *BigMatch* and *SASMatch* were able to successfully complete all of the tests, *fastLink* was unable to perform the CUF deduplication using the current blocking scheme in a reasonable time frame, and MAMBA did not complete any of the tests. Our results were written-up in an internal document, including user experiences that should be helpful for informing further testing with novel tools. These investigations have also assisted in understanding the Census Bureau's needs and how well the current offerings meet these needs to direct the future of record linkage at the Census Bureau.

Staff: Daniel Weinberg (x38854), Yves Thibaudeau, Chad Russell, David Brown (CES), Tom Mule (DSSD)

H. 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 FY2022, the Coverage Measurement staff was preparing for the public data release of the coverage estimates for Census 2020.

Staff: Jerry Maples (x32873), Ryan Janicki

I. 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 FY2022, staff began developing statistical models to predict block group reliability using Census 2010 PL94-171 demonstration data. Staff identified appropriate covariate data, including ACS data, and comparable Census 2000 tabulations. For the set of block groups composed of less than 300 persons, the statistical model is able to accurately predict reliability for

67% of block groups, which represents an improvement of 10% over the naïve benchmark of 57% reliability.

Staff: Kyle Irimata (x36465), Tommy Wright

J. 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 FY2022, staff developed side-by-side comparisons of Demonstration Data tables using the *Demographic and Housing Characteristics (DHC) File* version of the DAS against the published 2010 Census tables based on swapping (SWA). We compared county level data from Maryland, Delaware, Minnesota, and the District of Columbia.

In our FY2023 work, we will work to establish a potential list of DHC data that could possibly be used internally as a source of covariates to aid in potential use of statistical modeling in Census Bureau data and data products, especially to compensate for missing data due to non-response in Census Bureau sample surveys and censuses.

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

K. 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 FY2022, staff worked with stakeholders to produce statistical models for noisy measurements. A suite of evaluations was developed based on accuracy of predictions of true underlying counts and coverage/width of associated intervals. Several variations of models were compared to each other and to noisy measurements from the DAS. Staff implemented a method to create design matrices which associate previous tabulations - e.g., based on prior census counts and ACS estimates - to the noisy measurements; this method can be used even when table shells do not align. Several Bayesian hierarchical models were considered with the DAS noise mechanism approximated at the top level. This included a lognormal regression model and a multivariate spatial mixture model. The former provided a significant reduction in error from the noisy measurements for the majority of tabulations, while the latter provided coverage closer to the nominal level. To support future development work, staff developed documentation and tools in R, Markdown, and Knitr to generate deliverables and facilitate collaboration with stakeholders. Staff completed work on a manuscript discussing the use of the continuous Gaussian distribution as an approximation to Laplace or discrete Gaussian DP noise mechanisms in a Bayesian hierarchical model. Staff received reviewer feedback on a manuscript on spatial change-of-support modeling of noisy measurements and submitted a revision.

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

L. 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 FY2022, staff had regular conversations with Decennial Statistical Studies Division (DSSD) colleagues to understand the problem and project goal; read literature on simultaneous edit and imputation using Bayesian hierarchical models; provided advice on computational routines in Python for initial programming of the joint edit and imputation models described in Akande et al. (*JSSM*, 2019); and assisted with developing code for multiple imputation chained equations methods in R.

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

M. 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 FY2022, staff participated in regular meetings (both as a larger team, and in small groups) to discuss 2020 Census data and procedures conducted during that enumeration in order to gain insights on how to predict GQ counts. Staff are considering various statistical models and assessing their performance on 2020 GQ counts in nursing homes and dormitories to offer context regarding errors and outlier detection.

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

N. 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 FY2022, staff engaged with a decennial project team about analysis and simulation objectives, procured available datasets, and prepared computing environment. Initial data analysis efforts focused on tract-level response rates observed at the end of data collection in the 2020 Census. A spatio-temporal proximity metric was developed to quantify the exposure of a tract to MQA events using their locations and dates. Ideas for a simulation framework to compute household-level response probabilities (and related quantities) were discussed with the project team.

Staff: Andrew Raim (x37894), Lisa Moore (DCMD), Austin Schwoegl (DSSD)

O. 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 Survey (through a contractor Young & Rubicam) on attitudes to the decennial census and their relationship to completing the census. 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 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 FY2022, staff continued research into survey weighting methodology for calibration adjustment of weights applicable to the Tracking Survey. Staff wrote and published a long, detailed Technical Report (in the *CSRM Research Report Series*) summarizing new methodological developments and data results achieved on this project. The new methodology consisted of: (i) theory supporting large-sample consistency, asymptotic normality and variance estimation for estimates of survey attribute totals using poststratified weights; (ii) methods of handling missing calibration-variable data in base-weight creation and poststratification; and (iii) comparison of RDD and Web survey estimates, and assessment of survey weighting through agreement of benchmark proportions with known national targets. The methods were applied in the Report to the Tracking Survey data. This report supplies the technical basis for reports currently being written by the CBSM Tracking Survey Team on the survey's data quality and the comparison between its probability and nonprobability survey results.

Novel elements of this research concern design-based extensions of the calibration theory of Deville and Sarndal (1992) to the case where large weight-adjustments are necessary; a new approach to the handling of missing item-data in poststratification, without the need to impute simultaneously all missing calibration-variables; and new methods for the assessment of effectiveness of weight-adjustment, with application to the Tracking Survey. Staff (Slud) presented talks on the weighting-methodology research at the 2022 Small Area Estimation Meeting and the 2022 Joint Statistical Meetings.

The Team continued meetings aimed at writeup of final report on the Tracking Survey weighting, data-quality checks and comparison of Probability and Nonprobability survey results with national benchmarks. This report will have segments included in the overall Decennial report on the 2020 Tracking Survey. Authors of the Tracking Survey report discussed with staff (Slud) a quick follow-on project to provide weighted daily estimates on the Tracking Survey's attitude and census completion variables. This task involves new work on development and documentation of R code to re-weight the daily

observations using a moving window of month-long Phone and Web survey data to produce the daily outcome data. Re-weighting is need due to slight differences over time in the survey respondents' demographic characteristics.

Staff: Eric Slud (x34991), Darcy Morris, Jennifer Hunter Childs (CBSM), Casey Eggleston (CBSM), Jon Krosnick (CBSM).

P. Experiment on Minimum Height of Text Display on Mobile Survey

Description: An experiment was carried out to investigate the minimum height of text displayed on a smartphone screen so that the reader can read the text with ease. This was a one-factor between-subjects design. The experimental factor was the height with 3 levels (1 mm, 1.5 mm, and 2 mm). A participant was randomly assigned to one of the three levels and was asked to perform a reading task. The responses obtained were the number of errors per 100 words made during reading, time (in seconds) taken to finish the reading, and subjective rating of perceived difficulty level in performing the task, using a 5-point rating scale (very easy, easy, not easy nor difficult, difficult, very difficult).

An additional experiment was carried out to assess the most preferred x-height of a text that a respondent can read with ease. Five paragraphs in five different heights respectively were printed in black on a single letter-size white paper. The five heights were: 1 mm, 1.5 mm, 2 mm, 2.5 mm, 3 mm.

Highlights: During FY2022, the data analysis showed that there was no statistically significant difference in reading errors or reading time among the three height displays. The findings showed that a height of 2 mm was the most preferred height.

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

Q. Experiment on Optimal Font Style for Outdoor Viewing of Mobile Survey

Description: An experiment was carried out to investigate the font style of text displayed on a smartphone screen so that the reader can read the text with ease in shaded outdoor condition. The participants in the experiment performed a reading task in a shaded outdoor setting. This was a one-factor between-subjects design. The experimental factor was Font Style with 2 conditions (Regular, Bold). Each participant was randomly assigned to one of the two conditions. The responses obtained were the number of errors per 100 words made during reading, time (in seconds) taken to finish the reading, and subjective rating of perceived difficulty level in performing the task, using a 5-point rating scale (very easy, easy, not easy nor difficult, difficult, very difficult).

Highlights: During FY2022, the data analyses involved comparison of outcomes (error rate, reading time, and preference) among the two experimental conditions. The findings showed that, there was no statistically significant difference in reading errors or reading time between the two font styles. However, Bold font was preferred by many more participants.

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

R. Identifying Optimal Size of Touch Target for Mobile Survey Instruments

Description: In mobile survey instruments, the design of touch target has particular implications to survey data quality because responses to survey questions are often made through tapping one or multiple touch targets. A key concern over mobile survey instruments is the design of the smartphone screen (user interface) because it has a significant impact on the quality of data entered by survey respondents. The small screen size and touch interface of smartphones presents usability challenges affecting the effectiveness, efficiency, and satisfaction of respondents' interaction with mobile survey instruments. This brings up the question of the optimal size of a touch target that is small enough to best utilize limited screen real estate on a smartphone but large enough for successful data entry. The present study investigated touch-target tapping behavior by systematically manipulating the combination of touch target size and spacing. The following specific research questions were addressed: (1) What is the pattern of target-tapping success rate as a function of touch target size? (2) For a given target size, is the likelihood of successful target-tapping modulated by the spacing between the target and its surrounding objects? (3) What is the combination of minimum target size and spacing that yields a target-tapping success rate of at least 80%? (4) What is the relationship between touch target size and time taken to tap the target? (5) Are there any differences in tapping behaviors between a square touch target and a circular touch target?

Highlights: During FY2022, the experiment and the corresponding data analysis yielded the following conclusions and recommendations: (1) target touch rate increased with increase in target size, (2) target touch rate reached 80% at 6-mm target size and approached an asymptote of 100% at 11-mm, (3) target touch time decreased in small decrement with increase in target size, and (4) target spacing had no effect on tapping behavior. A 6-mm touch target for mobile survey response options is recommended.

A manuscript based on the above investigation has been submitted for publication.

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

S. Design Testing for Reporting Zero in a Web Survey
Description: When survey respondents do not have anything to report for questions asking for a quantity, it is important that they report their lack of quantity instead of skipping the question altogether. These reports of zero have historically been collected by asking respondents to check a "none" box. The effectiveness of this design has not been well supported, and with the growing prevalence of web surveys, new design options are available. An experiment was carried out to test the impact of three designs for reporting zero within a web survey on data quality and user friendliness: a "none" box, instructing respondents to enter "0" if they have nothing to report, and asking a yes/no filter question before asking for a quantity. An online questionnaire was administered to a panel of respondents across the U.S. who were randomly assigned to one of the three designs. Respondents completed the questionnaire using PCs and received four questions with expected high proportions of reporting zero. To account for repeated measures across each respondent, a hierarchical random effects model was used. Separate models were used to compare each of the data quality and user friendliness metrics across design conditions.

Highlights: During FY2022 and after controlling for demographic variables, it was found that question design did impact data quality and user friendliness. The "none" box design resulted in significantly more missingness and user errors than the enter "0" design, which also had significantly fewer errors than the filter question design. It was found that the "none" box design was, on average, the least efficient and was preferred the least by respondents. The enter "0" design, on average, had the fewest clicks, was the most efficient, and tended to be preferred over the "none" box design.

Staff: Thomas Mathew (x35337), Jonathan Katz (CBSM), Rachel Horwitz (DSMD)

T. 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 FY2022, the universities focused on technical work and reported their progress by giving seminars, written updates, and internal technical reports. Staff worked with and provided feedback to the universities.

Staff: Rebecca Steorts, Emmanuel Ben-David, Dan Weinberg, Krista Park (CODS), Anup Mathur (CODS)

1.14 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 6385H70)

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 FY2022, staff met with external researchers to propose methodology that provides privacy protection to ACS data, while maintaining its spatial correlation structure. Plans to form a consortium were discussed. Also, staff completed an initial draft documenting a research project on generating custom ACS estimates from a continuous-time model.

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 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 FY2022 and following delivery of the 2021 VRA estimates and variances in FY2021, further activity included: completed an Executive Summary of the project methodology for the public data release in December 2021; added to the documentation of the code base and planning a timeline the technical documentation

of the project to be completed by fall 2022; filled in some details of the technical documentation; and attended lessons-learned meetings with the Steering Committee to plan improved project management for the next (2023-26) cycle of VRA activity under different CSRM leadership.

Staff began regular meetings and preliminary reporting on VRA code-base and methodology looking ahead to the 2026 VRA cycle.

Staff conducted intensive analysis of data to create exhibits for the Final Technical Documentation of the 2021 Voting Rights Act project, and extensive writing of the Final Report for the public-facing VRA website and also for the *CSRM Research Report Series*. The document will be delivered in mid-October 2022 for review.

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

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

Description: Census Bureau 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). 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 FY2022, staff had regular conversations with Decennial Statistical Studies Division (DSSD) colleagues to understand the problem, the experimental weighting methodology, and related methodologies used in other Census Bureau sample surveys; performed a code review of the experimental weighting methodology and documented findings; and started working with a research dataset of response status and administrative record variables for 2018-2020 ACS sampled units. With this dataset, staff looked at monthly comparisons of respondent vs. non-respondent characteristics, began developing response propensity models for ACS 2020 data to investigate the influences of covariates on the response status, and identified machine learning models (lasso, random forest, boosting) for response propensity that deal with high dimensionality and accommodate possible complex interactions. These model developments will ultimately lead to response probability estimates that can be 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. Properties and performance of this method was assessed in a simulation study as described in Section 2. By the end of FY2022, staff had formulated a study plan for a two-stage weighting approach that involves response propensity modeling and calibration and began documenting gradient boosting machines (GBM) results and assessing performance.

Staff: Darcy Steeg Morris (x33989), Joseph Kang, Patrick Joyce, Isaac Dompreh, 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 FY2022, staff continued with simulations to explore survey characteristics associated with SDR biases in small domains. Aspects of PPS sampling design along with skewness and kurtosis of the distributions of outcome variables were found to have some effect on the bias of SDR estimates. An unexpected finding was that the biases could also depend on the sizes of domains in the same survey other than the domain of interest, due to the 'recycling' pattern of SDR replicates.

Further staff activity: theoretical write-ups concerning the biases of SDR and Successive-Difference variance estimation methods in independent identically distributed superpopulations; simulation runs and interpretations

under stratified random sampling and PPSWOR sampling designs; and further checks of the simulation code, in preparation for the final set of runs to be compiled into the projected simulation study as a *CSRM Research Report Series* and journal-paper submission.

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

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

A. Data Integration

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

Highlights: During FY2022*,* staff has been documenting Record Linkage and unduplication code in preparation to release code. A new source-documented BigMatch is ready for release. Staff started a project to attack the Current Population Survey Public Use Microdata Files. The project was defined as follows. Staff will attack the files using administrative files and other files available to the general public. Staff will identify households which are believed to be in the Current Population Survey. That list will be provided to a member of the Demographic Surveys Division staff who will report back the number of correct and incorrect identifications. Staff will have no knowledge on which method of disclosure avoidance was used for the CPS files.

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

1.17 POPULATION DIVISION PROJECTS (Demographic Project TBA)

1.18 SOCIAL, ECONOMIC, & HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165022)

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 FY2022, staff attempted multiple approaches to model correlated shares of poor and nonpoor children in school district pieces within county to test whether the independence assumption between the sets of shares held. None of the approaches, both frequentist and Bayesian, yields workable models. Work on this line of research has been terminated due to lack of feasible model frameworks.

Staff has started evaluation of the Dirichlet-Multinomial Share model for school district estimates. Given the large number of school district and counties, the assumption that the model parameters are the same across all areas is very strong. Staff applied model diagnostic techniques to examine whether some subsets of the data behaved differently from the rest. Staff focused on county population size and county-level income tax geocoding rates to partition the data into groups. The key findings were that the model precision parameter was increasing for groups of counties with larger population and for groups of counties with high geocoding rates. This resulted in estimates with larger predictive MSE on average (especially for the school districts in smaller/low geocoding rate counties). Further evaluation is needed to understand the impact to the point estimates for school district shares for official use. Results from this project were presented at the Small Area Conference at University of Maryland, College Park on May 26, 2022.

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 FY2022, staff generalized the ideas behind the nonparametric regression method called Geographically Weighted Regressions (GWR) to incorporate into the Fay-Herriot Small Area model framework. The GWR allow for parameters to be

referenced spatially and are fitting using only the data in a local neighborhood. This is achieved by down weighting observations that are not 'near' to the reference point. Staff has developed the algorithms to perform the weighted regressions, predictions and MSE calculation within the Fay-Herriot model framework. Staff also introduced a new bandwidth selection criteria which aligned with the goals of SAE (reduction in predictive MSE).

For demonstration purposes, staff created a public use dataset from the 5-year 2015 ACS released estimates for county-level poverty (all ages) and the number of people by county receiving Supplemental Nutrition Assistance Program (SNAP) benefits for 2015. Staff applied the weighted regression techniques developed on the poverty dataset. Staff was investigating whether the parameters in the Fay-Herriot model to predict the log of the number of people in poverty at the county level varied as a function county population size. Staff has generalized the idea of geography in the GWR framework to a more general space defined by a set of factors, e.g., single factor such as county size, geographic coordinates, multivariate vector of characteristics, etc. Staff analyzed the demonstration dataset using three different definitions of space: county size (univariate factor), geographical closeness (usual geography) and multivariate vector of percentage of minority population, percentage of college educated population (25+) and percent of employed population (18+). The space defined by county size showed differences in the intercept terms across the areas. No differences were detected among the parameters using geography and only a slight difference in the intercept term for the multivariate factor of characteristics. Results from this work were presented at the Joint Statistical Meetings held in Washington, D.C. in August 2022.

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: In FY2022, staff continued development of nonparametric multivariate spatial small area models for unit-level survey data which can be applied to estimation of the number and proportion of the population with health insurance coverage. Code was written for efficiently fitting this class of models to large data sets using different families of pseudo-likelihoods, and these functions were compiled into an R package. Staff continued work on a paper documenting this methodology and code.

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

1.19 GENERAL ECONOMIC STATISTICAL SUPPORT (Economic Project 1183X01)

1.20 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 FY2022, staff continued working with Economic Statistical Methods Division and Research & Methodology Directorate staff on 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. Staff received internal review feedback and edited paper accordingly.

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 FY2022, staff provided seasonal adjustment and software support for users within and outside the Census Bureau and completed a document that provides recommendations on assessing residual seasonality in economic time series. Staff gave specific seasonal adjustment support to Instituto Brasileiro de Geografia e Estatística, Fidelity Investments, Ecole Nationale Superieur de Statistique and d'Economie Appliquee, Statistics Austria, The Better Policy Project,

Frantz Quantitative Research Proprietary Labs, UTokyo Economic Consulting, Balyasny Europe Asset Management, ExodusPoint Capital Management, Bank of Nova Scotia, and California Department of Finance. Staff continued to update the GitHub repository for Ecce Signum.

Staff: Tucker McElroy (240-695-3610; R&M), James Livsey, Osbert Pang, William R. 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 2020 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 FY2022, Build 59 of X-13ARIMA-SEATS was sent to the Economic Directorate for internal testing before public release. Development of a python program to run X-13ARIMA-SEATS continues to be developed.

The new build (Build 60) of X-13ARIMA-SEATS was released for internal testing with internal stakeholders. Discussion continues about functionality of new python seasonal adjustment platform.

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

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 FY2022, staff made progress on several research projects: (a) completed documentation for the assessment of weather effects on seasonal adjustment; (b) conducted new simulations and assessment of benchmarking optimization methodology for indirect seasonal adjustment of quarterly time series; (c) polished final results on the use of the EM algorithm for fitting multivariate time series; (d) applied maximum entropy outlier framework to time series affected by Covid-19; (e) developed principal components methodology for an index of economic activity, based on Census time series. Also developed nowcasting methodology to make the index publication more timely; (f) applied benchmarking techniques to constituent variables of GDP, in order to remove residual seasonality while preserving aggregation relations. Developed hierarchical lattice structure of aggregation relations and extended benchmarking code into a top-down algorithm. (g) fitted time series models to 12 major sub-aggregates of GDP and extracted business cycle estimates in a study of how cycles may help explain movements in GDP. Then we assessed correlation structure and Granger causality of extracted cycles, and compared forecasting results from a large class of regression models; (h) developed and fitted weekly time series models that take account of the non-integer period of seasonality; (i) studied multivariate model-based seasonal adjustment applied to manufacturing time series; (j) revised writing and simulations for a study of mean squared error of seasonal

adjustments from differing frameworks; (k) devised code and methodology to automatically identify temporary change regressors.

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

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 FY2022, X-13ARIMA-SEATS manual was updated, including the text for the -c flag (running X-13 via command line) to reflect that the current behavior is to suppress the indirect adjustment in a composite run.

The Census Bureau decommissioned the legacy census.gov pages, and staff made all relevant changes to existing documentation to support all links and references. Additionally, the X-13 manual was updated with the most recent references.

Finally, staff updated an example in the manual for the estimate spec. Additionally, staff included new documentation for save tables in the SEATS spec of the X-13ARIMA-SEATS reference manual.

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 FY2022, M3 aggregate series were considered for investigation. Here, lower-level series are seasonally adjusted using X-13ARIMA-SEATS as well as adjusted jointly using a multivariate method for the purpose of making empirical comparisons of indirect seasonal adjustment comparisons of the aggregate series. Currently forecast error and revision history for the two separate methods of indirect seasonal adjustment are being investigated.

Staff devised and wrote code for a simulation study on the effects of multivariate modeling under multiple scenarios. These scenarios include a range of cases that include when trends and seasonal components are highly correlated or uncorrelated. Multiple data generating process models were coded and structural component models fit.

Staff began investigation into the effects of model misspecification and its effect on structural component modeling. Specifically, how model misspecification can influence multivariate seasonal adjustment. The initial findings are that when models are correctly specified there is substantial improvement when moving from univariate seasonal adjustment to multivariate seasonal adjustment.

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

G. Classification of Businesses for the North American Industry Classification System (NAICS)

Description: This is an exploratory Investigation of data and methods to use machine learning approaches such as text mining techniques to automatically classify business establishments from different sources/frames according to the North American Industry Classification System (NAICS). Two such recent studies are (1) "Using Public Data to Generate Industrial Classification Codes" and (2) "Using Machine Learning to Assign North American Industry Classification System Codes to Establishments Based on Business Description Write-Ins."

In the first study, the investigators initially collected 1,272,000 records of establishments via a grid search on both Yelp and Google Places APIs, based on a combination of geo-coordinates and keywords in the titles of all two-digit NAICS sectors. Records that did not have a website and user reviews are eliminated reducing the collection to approximately 290,000 records. Training and evaluating models for classification purposes require a random sample of business establishments for which their NAICS codes are known. Next, the 290,000 records are then linked to establishments on the Business Register (BR) using the Multiple Algorithm Matching for Better Analytics (MAMBA), a fuzzy matching software developed by Cuffe and Goldschlag. Linkage and other restrictions imposed on selections resulted in a final collection of records for 120,000 single-unit

establishments. Employing doc2vec, a text mining technique, the textual information in each record is transformed to vectors. These vectors and series of binary variables indicating the Google Type**,** tags are used as features, i.e., predictors of the NAICS codes. In this approach, Random Forest models are trained to predict NAICS codes. It is reported that the best model performs approximately 59% accurately. Overall, this work initiates an interesting approach for the NAICS classification problem, but one main question is to what extent the methodology can be relied on? One main issue is selection bias and the non-probability nature of the collected data. The data collection process seems to systematically exclude and include business establishments. For example, due to coverage in the source of the collection, grid search selection mechanism and importantly**,** the error-prone record linkage. A closer examination of the data may shed light on these issues.

In the second study**,** the NAICS codes are assigned to business establishments in the Economic Census. Our understanding of this work is based on less detailed information from the presentation given at the 2019 Joint Statistical Meetings. In this work, self-designated kind of business write-ins from the 2012 Economic Census**,** textual information in combination with business names and line labels are used to predict NAICS codes. The textual information is transformed to vectors using the bag of words approach. Two classification methods employed here are Naïve Bayes and Logistic regression. These models are trained on 339,936 records. The performance of each selected model is tested on 37,772 records. In the presentation, it is reported that Logistic regression using write-ins, business name, and line label as their features for predicting NAICS codes performs the best. Naïve Bayes and Logistic regression are very basic classification methods and more advanced classification methods can improve the results and change the findings. Also**,** doc2vector transformation provides more effective representation of text than that of bag of words.

Highlights: During FY2022*,* there was no progress on this project. We are awaiting to resume activities after training some new members.

Staff: Emanuel Ben-David (x37275), A.J. Goldsman (Deloitte), Javier Miranda (Halle Institute for Economic Research), Ann Sigda Russell (EWD), Andrew Naviasky (EWD)

H. Production and Dissemination of Economic Indicators

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

Highlights: During FY2022, staff wrote and revised code to estimate price indices using two data sources: Nielsen retail scanner data and NPD Group data. Staff discovered a coding bug that affected the estimation of the CUPI and SUPI price indices from the Nielsen data. After correcting the bug, staff began to recalculate the full results with the help of co-authors at the University of Michigan. Staff re-wrote code to estimate CUPI and SUPI price indices on the Integrated Research Environment (IRE) in Python, adding several new features that did not exist in the original R code base. Using this code and data from The NPD Group, staff estimated CUPI and SUPI price indices for three types of products (coffeemakers, headphones, and memory cards) at the state level within the continental United States. Staff used the indices for these product types to produce aggregate indices reflecting the expenditure weighted average cost of living across all three goods. Staff presented visualizations of these results at JSM. Staff will seek support to overcome barriers to make these methods viable for application.

Staff: Adam Hall (x32936)

I. 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 FY2022*,* the working group discussed several proposals for improving the LRS. We discussed alternative data sets for predicting the LRS. A few data sets were identified and transferred to IRE environment. We proposed different approaches for improving the LRS model including spatial regression, fix and random effect models. Another comprehensive study undertaken was the measure of uncertainty and margin error for each predicted LRS.

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)

1.21 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 FY2022, staff continued to maintain and support the Integrated Research Environment (IRE) and the Cloud Research Environment (CRE) prototype. With the Center for Optimization Data Science, Computer Services Division, and Chief Technology Office, we 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.). We built out the initial set of servers of IRE New, and these are now undergoing user acceptance testing. We expect the migration to be complete by FY2023/Q2. In CRE, several new projects were provisioned, including a project to analyze the rates of PIK assignment for the 1940 Census, and two projects to analyze record linkage software with various types of input files (demographic versus business data).

Staff: Chad Russell (x33215)

1.22 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 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 FY2022, staff performed weighted Regression Analysis and Bayesian hierarchical modeling to produce county-level direct survey-based estimates for thirteen tobacco smoking outcomes (e.g.: Current smoking prevalence among age 18+; Ever smoking prevalence among age 18+; Percentage of people whose workplace does not allow smoking (among age 18+); Percentage of people who live in a residence where smoking is not allowed (among age 18+); etc.). County-level designed-based estimates for these tobacco outcomes were calculated for 3,134 counties across the country. Model-based estimates for population coverages for these 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 calculated posterior probabilities for each of the thirteen smoking outcomes.

Staff conducted model diagnostic tests to validate the final Hierarchical Bayesian (HB) estimates. Staff used the model diagnostics to select the best models based on the Arcsine transformation methods. Staff is preparing final delivery to NCI Final estimates to NCI must go through the U.S. Census Bureau Disclosure Review Board first. Final validated county level HB estimates will be benchmarked to the corresponding state level direct estimates.

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 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 FY2022, 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 participated in a series of lectures given by Jun Shao and informal presentations on missing data problems at the Census Bureau, studied historic and recent literature on calibration techniques for nonresponse.

Staff continued hands-on learning of such methods through assessments of the ACS experimental weighting procedure for 2020 ACS data products and working groups sharing ideas for alternate methodologies. Motivated by this application, staff documented a study plan for the study of a two-stage (inverse probability followed by calibration) approach for adjusting for dynamic and large magnitude unit missing data. This study will develop 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 conferences. Staff worked on a manuscript of this work to be submitted to a peer-reviewed journal.

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, 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 FY2022, staff worked with Economic Statistical Methods Division (ESMD) staff on 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: Darcy Steeg Morris (x33989), Joseph Kang, Isaac Dompreh, 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 FY2022, staff worked to organize and provide a series of four lectures (90 minutes each) that provided introductory concepts, definitions, theory, and applications of statistical methodology. Topics covered include—missing data mechanisms (missing completely at random, missing at random, covariate-dependent missingness, and nonignorable missingness) and ignorable missing data (observed likelihood and MLE under missing at random, semi– and non-parametric methods under covariate-dependent missingness, imputation methodology, and variance estimation). With the help of Division Chiefs, participants (in parenthesis) were identified across the Census Bureau as follows: [Center for Economic Studies (1), Center for Statistical Research & Methodology (6), Demographic Statistical Methods Division (4), Decennial Statistical Studies Division (6), Economic Statistical Methods Division (4), Research & Methodology Directorate (1), and Social, Economic, & Housing Statistics Division (1)].

Based on feedback from all 23 participants and discussions, the participants agreed on a follow-up to the lectures would consist of a series of presentations under the title "Missing Data; Problems, Applications, & Software Presentations/Lectures." Scheduled discussion leaders and topics are: (1) Joe Schafer (May 20, 2022, some recent work related to the American Community Survey); (2) Jonathan Rothbaum (June 17, 2022, some recent work on the Current Population Survey); (3) Joseph Kang (September 23, 2022, some ideas on building sample survey weights using a machine learning method and the entropy balancing calibration technique); and (4) Mark Asiala (October 21, 2022, some recent work on the American Community Survey).

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 FY2022, staff studied a few new scenarios that involve the general problem of regression with linked data sets, including survival data analysis with Cox model, and Small Area Estimations. Two working papers on this research topic are currently under development. Staff submitted the paper "Regularization for Shuffled Data Problems via Exponential Family Priors on the Permutation Group" to a machine learning journal. In this paper, we propose a flexible exponential family prior on mismatches that can be used to integrate various structures such as sparse and locally constrained mismatches. This prior turns out to be conjugate for canonical shuffled data problems in which the likelihood conditional on a fixed permutation can be expressed as product over the corresponding (X, Y)-pairs. Inference is based on the EM algorithm in which the intractable E-step is approximated by the Fisher-Yates algorithm. Staff published the second paper "Regression with linked datasets subject to linkage error" in Wiley Interdisciplinary Reviews. In this focus paper, staff members give an overview of the current literature on the problem of regression with linked data files, with an emphasis on recent approaches and their connection to the so-called "Broken Sample" problem. Staff members also provide a short empirical study that illustrates the efficacy of corrective methods in different scenarios.

Staff: Emanuel Ben-David (x37275), Guoqing Diao (GWU), Martin Slawski (GMU), Zhenbang Wang (GMU)

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 FY2022, Reddy and Steorts performed comparisons using Sadinle (2018) and Edmorando et al. (2018) on real and synthetic data, where open-source code and a paper were developed. Staff continued record linkage comparisons, received reviews on their paper and are working to revise the reviewer's comments.

Specificallly, staff compared popular methods in the literature, such as Fellegi and Sunter (1969), fastlink, and the approach of Sadinle (2014) on both real and simulated data to gain an understanding regarding how Fellegi-Sunter type models perform in terms of record linkage evaluation metrics in the literature. We do not consider other methods as our goal is to only consider those that follow the Fellegi-Sunter framework in the paper. The code for all the models is reproducible and the paper is currently under revision. Our hope is that this may provide guidance moving forward regarding the Census Bureau regarding what types of methods may be helpful regarding these types of methods moving forward or provide guidance for future extensions.

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 FY2022, staff finalized a review paper on record linkage/entity resolution, which is in press at Science Advances. Staff gave talks and tutorials.

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

D. Analysis of Alternatives and Proof of Concept for Record Linkage Modernization

Description: The Census Bureau is a pioneer and has a long tradition using record-linkage methodology for multiple endeavors, such as unduplication of administrative lists and census post-enumeration studies. The Census Bureau also link administrative files to support many research projects sponsored by non-profit and academic institutions. Some of the record-linkage systems at the Census Bureau are decades old and there have been only a few upgrades in the basic Census Bureau record-linkage methodology over that time, barring few exceptions. There are several stand-alone software packages in operation even as there is scant coordination or integration of these packages at the enterprise level. At the same time the computing capabilities available in modern ecosystems are currently under-exploited. The Census Bureau is embarking in a comprehensive modernization and software engineering effort to overhaul and integrate the legacy record-linkage infrastructure at the Census Bureau along with new technology for enterprise-wide functionality. The Center for Statistical Research and Methodology (CSRM) is being called upon to work with senior computer scientists at the Census Bureau and software engineering professionals to elicit industry-grade requirements directed at identifying the record-linkage solution that best serves the need of the Census Bureau. The technical knowledge and institutional memory of CSRM are crucial in supporting the requirement process that will be provided to vendors and developers so they can design and provide a record-linkage solution that integrate the most recent breakthroughs in record linkage and is fully usable in the context the multiple applications of record-linkage at the Census Bureau.

The Center for Optimization and Data Science (CODS) and CSRM are responsible for the execution of a "proof of concept" which involves narrowing a pool of dozens of candidate vendors and open source solutions down to 5 or 6 finalists. These finalists will then be the object of extensive testing. Tests will involve simulations as well large benchmark record-linkage exercises, such as unduplicating the Decennial Census, to evaluate the accuracy, performance, and usability of the prospective solutions. CSRM expects run over 30 specific tests for each candidate for a total of approximately 200 runs.

Highlights: During FY2022, staff completed Quantitative and Qualitative Investigations (QaQi): Case studies were designed specifically QaQi. A decennial case study and a business case study. Both case studies had multiple parts. With the collaboration of CODS along with staff of the Decennial and Economic areas staff managed the case studies. Staff played an important role in setting up the "truth decks" that were the lynchpin of the investigation. Staff developed the scoring methodology centered on a dynamic "F1" score. The F1 score reflects the performance of a particular record-linkage algorithm, and it evolves as the cutoff between alleged matches and nonmatches is dragged down the scale of Fellegi Sunter weight and/or of the posterior probabilities. Staff showed that a matcher will capture information on the true status of the pairs if the associated F1 curve peaks before fading. This criterion allowed staff to compare the discriminating power of the algorithms in identifying "true matches." A highly discriminating algorithm exhibits a monotone F1 curve that will reach its peak early and a value close to 1 before quickly dropping. The flatter the F1 curve is, the, the poorer the discrimination power of the algorithm is. Staff plans to continue researching performances of record-linkage algorithm and methods to best quantify and compare them.

Staff: Yves Thibaudeau (x31706), Daniel Weinberg, Chad Russell, Jaya Damineni (CODS), Yatish Koli (CODS), Anup Mather (CODS), Steve Nesbitt (CTR)

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. Household Survey Design and Estimation
[See Demographic, Economic, and ACS Projects]

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 FY2022, staff finalized a visualization for visualizing uncertainty in an estimated overall ranking of K different populations based on Klein, Wright, and Wieczorek (2020) and software (Wieczorek, CRAN). As with earlier work on comparisons reported during FY2021, staff provided estimated overall rankings of the states and a joint confidence region for 88+ different American Community Survey (ACS) topics for each of the years 2018 and 2019. As with comparisons, underlying methodology adjusts the level of significance assuming independence only.

Before finalization, staff obtained extensive individual feedback from almost 20 nonrandomly selected Census Bureau employees on the two visuals (1) Comparisons and (2) Rankings with Tutorial. The extensive feedback covered a variety of topics. Staff used the extensive feedback to revise the visuals and reached out to the Center for New Media & Promotion and the Applications Development & Services Division and posted the 2 visualizations as parts of The Ranking Project on our center'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 FY2022, staff worked to develop a framework for applying optimal exact sample allocation methodology to the overall ranking methodology discussed in the Ranking Project in an effort to optimally tighten the joint confidence region, either for an overall fixed sample size for all of the populations being ranked or for a sample size that results when the sample is sequentially allocated among the populations being ranked. A few properties were given which hold the key to tightening a joint confidence, all pointing to the number of overlapping joint confidence intervals.

Staff: Tommy Wright (x31702)

Apportionment
Highlights: During FY2022, staff worked on a manuscript for apportionment with a focus on the Method of Equal Proportions including an explanation of the Hill methodology and the need for the Huntington methodology which corrected a flaw in Hill's procedure. This is done in the context of Lagrange's Identity. Staff also published a related paper Wright (2021).

Staff: Tommy Wright (x31702)

D. Consistent Estimation of Mixed-Effect Superpopulation-Model Parameters in Complex Surveys with Informative Sampling
Description: This research studies the problem of design-consistent model-assisted estimation for regression and variance-component parameters within parametric models based on complex survey data. Starting from seminal work of Binder (1983) on 'pseudo-likelihood,' it has been known how to design- and model- consistent inference from survey data, based only on observed data and single-inclusion weights, when units are independent under the superpopulation model. However, it has largely been an open problem since first studied in papers of Pfeffermann et al. (1998), Korn and Graubard (2003) and Rabe-Hesketh and Skrondal (2006), how – or if it is even possible -- to do consistent survey-weighted inference based on single-inclusion weighted survey data when data share random effects within clusters and sampling may be informative.

Highlights: During FY2022, there was no significant progress on this project.

Staff: Eric Slud (x34991)

E. MPPS Sampling in USDA Surveys
Description: This collaborative project explored the properties and sampling characteristics of MPPS (Multivariate Probability Proportional to Size) sampling, a Poisson sampling design used in FDA farm surveys with weights designed for inclusion properties with respect to multiple distinct crops.

Highlights: During FY2022, staff met regularly, resulting in preparation of a presentation by Dr. Cheng in the 2021 FCSM Research and Policy Conference on November 2, 2021. A paper based on the presentation will be prepared and submitted with USDA co-authors.

Staff: Eric Slud (x34991), Yang Cheng (USDA-NASS), Lu Chen (USDA-NASS)

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 FY2022, staff completed development of bootstrap and analytic estimators of the MSE, without the assumption of normality of the random effects. Staff proposed a new three-point mixture distribution, which matches the estimated moments of the random effects. Bootstrap estimators of MSE using this moment-matching distribution had less bias as compared to traditional estimators, especially in the case of a small to moderate number of small areas, or when the sampling variance is large, relative to the random effects variance. Staff developed methodology for estimating the MSE of unsampled (or unpublished) small areas for these analytic and bootstrap approaches and discussed similar adjustments to existing methods. Staff applied these approaches to county-level ACS data in Maryland and in Georgia, and also provided an example of estimating MSE for unpublished 1-year estimates in Georgia. Staff presented results from this research at the 2022 Small Area Estimation Conference and the 2022 Joint Statistical Meetings 2022.

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

B. Small Area Estimation for Misspecified Models
Description: Model-based methods play a key role to produce reliable estimates of small area means. These methods facilitate borrowing information from appropriate explanatory variables for predicting the small area means of a response variable. In the frequentist approach the empirical best linear unbiased predictors (EBLUPs) of small area means are derived under the assumption of a true linear mixed-effects model. Under the assumed model, these are approximately best predictors of the small area means. Accuracy of the EBLUPs are evaluated based on approximate mean squared error (MSE) of the EBLUPs, assuming the true model holds. Second-order accurate approximation of the MSE and its estimation, where all lower order terms are ignored in the asymptotic derivation, are the main objects in small area estimation.

Highlights: During FY2022, there was no significant progress on this project.

Staff: Gauri Datta (x33426), Eric Slud

C. 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 FY2022, the staff with an external collaborator completed a manuscript which is now accepted for publication in *Survey Methodology*.

Staff: Gauri Datta (x33426), Ryan Janicki, Jerry Maples

D. 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).

Highlights: During FY2022, based on the aforementioned project, staff submitted a manuscript, co-authored jointly with his two collaborators. A revised version of this manuscript has been submitted to a journal.

Staff: Gauri Datta (x33426)

E. 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 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. 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 FY2022, staff worked on the construction of an approximate highest posterior density (HPD) joint credible set of the population means. This credible interval is approximately elliptical. We simulated samples for the population means from this approximate 90% joint HPD credible region and used those samples to create a joint posterior distribution of the ranks of the population means. This approximate elliptical credible region turns out to be quite superior to the rectangular-based Bonferroni joint credible regions. More details will be provided in a manuscript.

Staff: Gauri Datta (x33426)

F. Machine Learning-assisted Fay-Herriot Model
Description: The Fay-Herriot (FH) model has been extensively used in small area estimation for official federal statistics with a clear goal of predicting a response of interest. Most of the FH-based models assume the linear combination of covariates and their coefficients, and hence they inherit problems in linear regression models: multicollinearity, nonlinearity, and complex interactions among the covariates. Some machine learning methods, including, the regression tree model, have overcome such issues to some degree. The staff has developed a tree-assisted 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 FY2022, staff built the tree-assisted FH model with a simulation study which showed that the tree-assisted FH model results in a more efficient variance estimate than the conventional FH model in the presence of strong interactions among the covariates while their main effects are less strong. This research was presented at the 2022 Small Area Estimation Conference. Based on feedback, we reached out to an external collaborator experienced in machine learning methods. We expanded the scope of this project to accommodate more machine-learning methods. Currently, we are implementing a cross-validated score into our algorithm to compare this with the tree-based method.

Staff: Joseph Kang (x32467)

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 FY2022, 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; (b) added changing frequency time series examples to a project that uses wavelets and the singular value decomposition to identify and extract seasonality; (c) finalized numerical work for a Bayesian method to identify and extract extreme effects through latent outlier processes; (d) developed new theoretical results for a mechanistic description of seasonality involving a marked point process.

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 FY2022, staff made progress on several projects: (a) refined and published the methodology for model identification based on testing for zeroes in nonparametric estimates of the spectral density; (b) devised a new method for comparing two specifications of differencing operator via multi-step ahead forecast mean squared error; (c) derived a new expression for the quadratic forecast filter that allows for recursive generation of nonlinear processes; (d) refined asymptotic results for polyspectral means, a type of estimator that involves a weighted integral of the polyspectral density; (e) revised results on maximal benefit of quadratic prediction over linear prediction; (f) refined simulations of new procedure for local spectral density estimation that is optimal at the boundary of the frequency domain; (g) refined work on time series differential privacy, involving new notions of utility and privacy, and the use of all-pass filtering as a protective device; (h) refined results on a new expression for the multivariate missing value filter, for the case of multiple missing values; (i) revised writing of a description of a new stable parameterization of GARCH processes.

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

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 FY2022, staff (1) developed a bivariate CMP distribution based on the trivariate reduction method, (2) developed a CMP-based longitudinal model, and (3) made further advancements to the COMPoissonReg R package. They wrote manuscripts associated with each of the respective projects above; Manuscript 1 was published in *Communications in Statistics - Theory and Methods* while Manuscript 2 was published in *Stats*. Manuscript 3 is being developed as a technical report. Meanwhile, staff are working to develop a CMP-based lag model that will allow for dispersion in observed count data.

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

B. Design and Analysis of Embedded Experiments
Description: The project's goal is to cover a number of initiatives based on the design and analysis of embedded experiments. Experiments carried out by the Census Bureau may occur in a laboratory setting but are often embedded within data collection operations carried out by the agency. Some organizational constraints require special consideration in the design and analysis of such experiments to obtain correct inference. Relevant issues include incorporation of the sampling design, determination of an adequate sample size, and application of recent work on randomization-based causal inference for complex experiments.

Highlights: During FY2022, staff completed a manuscript on the development of statistical methods to compare pairs of studies and determine whether one mailing strategy yields a more uniform call distribution than another in the distribution of call volumes to the Census Bureau's telephone helplines. The manuscript is now accepted for publication in the *Journal of Official Statistics.*

Staff: Thomas Mathew (x35337), Andrew Raim, Kimberly Sellers

C. Predicting Survey/Census Response Rates
Description: In this research, we study statistical models for accurately predicting U.S. Census self-response for identifying hard-to-count populations for surveys. The goal is to build models that allow for: interpretability without losing in predictive performance to state- of-the-art black-box machine learning methods, automatic variable selection in high-dimensional regression, and actionable interpretability for various levels of geography.

Highlights: During FY2022, a paper on this topic titled "Predicting Census Survey Response Rates via Interpretable Nonparametric Additive Models with Structured Interactions" was revised and resubmitted to an applied statistical journal. In this paper, we focus on interpretable models for accurate prediction of survey response rates. The U.S. Census Bureau's well-known ROAM application uses principled statistical models trained on the U.S. Census Bureau Planning Database data to identify hard-to-survey areas. An earlier crowdsourcing competition revealed that an ensemble of regression trees led to the best performance in predicting survey response rates; however, the corresponding models could not be adopted for the intended application due to limited interpretability. In this paper, we present new interpretable statistical methods to predict, with high accuracy, response rates in surveys. We study sparse nonparametric additive models with pairwise interactions via $\ell 0$-regularization, as well as hierarchically structured variants that provide enhanced interpretability. Despite strong methodological underpinnings, such models can be computationally challenging - we present new scalable algorithms for learning these models. We also establish novel non-asymptotic error bounds for the proposed estimators. Experiments based on the U.S. Census Bureau Planning Database demonstrate that our methods lead to high-quality predictive models that permit actionable interpretability for different segments of the population. Our methods provide significant gains in interpretability without losing in predictive performance to state-of-the-art black-box machine learning methods based on gradient

boosting and feedforward neural networks. Our code implementation in python is available at https://github.com/ShibalIbrahim/Additive-Models-with-Structured-Interactions. We are revising the paper on this topic that was submitted for publication.

Staff: Emanuel Ben-David (x37275), Ibrahim Shibal (MIT), Rahul Mazumder (MIT), Peter Radchenko (University of Sydney)

D. 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 FY2022, staff continued the preparation of a manuscript that reviews the literature on regression adjustment, covariate balance and re-randomization, with the goal of presenting a comprehensive review of the topics. Staff is also working on deriving an alternative to the difference-in-means estimator by utilizing the inclusion probabilities under the re-randomization criterion and computing the Horvitz-Thompson estimator. Progress has been made on the derivation of the first order and second order inclusion probabilities. The variance of the Horvitz-Thompson estimator will be derived, and its estimation will be addressed. In addition, the Horvitz-Thompson estimator will be compared to the difference-in-means estimator. It is anticipated that the proposed research will eventually lead to methodologies that can be applied to the analysis of some of the embedded experiments carried out at the Census Bureau. It is hoped that some datasets from ACS will be available for illustration.

Staff is currently exploring potential applications of the methodologies to some problems relevant to the Census Bureau. In particular, staff is looking at some split-panel experiments that have been carried out at the Bureau to see the extent of variance reduction that is possible under re-randomization. The specific experiment that is being considered addresses whether the day of the week a mailed survey invitation arrives at a housing unit affects the response rate to an online survey.

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

E. Statistical Properties of Differentially Private Observations Grouped into Intervals

Description: The project considers observations (such as income) grouped into intervals, and the counts in each interval are to be privacy protected before they can be released. For this, it is proposed to calculate the counts based on a sanitized version of the original data, after adding independently generated Laplace noise. The goal of the project is to investigate the statistical properties of the resulting counts, with a view to providing guidance on the choice of the privacy-loss budget. This will be contrasted with differential privacy protection applied directly to the counts and examining the statistical properties of the sanitized counts. The proposed research is expected to lead to methodologies that can be used to assess the statistical properties of sanitized data resulting from applying the TopDown Algorithm at the Census Bureau.

Highlights: During FY2022, there was no significant progress on this project.

Staff: Bimal K. Sinha, Kyle Irimata, Thomas Mathew

F. 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 FY2022, staff considered regression models to analyze data prepared using the differentially private mechanism Sufficient Statistic Perturbation (SSP). The proposed direct sampler was successfully implemented as a step within a Gibbs sampler to estimate parameters from the model. This work was presented at the AISC 2021 meeting. Staff prepared a separate manuscript on direct sampling with step functions; the manuscript features several applications - without differential privacy - which may be more familiar to a general audience of statisticians. After receiving an initial round of feedback, a revision of the manuscript was submitted. Reference implementations of the proposed sampler were prepared in Julia and R on Github.

Staff: Andrew Raim (x37894)

G. 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 FY2022, staff established methodology and implemented a prototype of the sampling framework in Julia. Applications to several low-dimensional toy problems demonstrated that samples conform to the target distribution. Initial ad hoc strategies were shown to effectively refine the proposal at lower dimensions.

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. Development and Evaluation of Methodology for Statistical Disclosure Control

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

Highlights: During FY2022, staff conducted research on statistical methods for protecting data confidentiality and respondent's privacy. Staff published a paper titled "A Review of Rigorous Randomized Response Methods for Protecting Respondent's Privacy and Data

Confidentiality" in a book edited in honor of Prof. C.R. Rao celebrating his 100th birthday. Staff worked on developing principled measures of the trade-offs between data confidentiality protection and data-utility loss and building optimal data perturbation algorithms.

Staff: Tapan Nayak (x35191)

B. 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 FY2022, staff worked on the paper "Bayesian Analysis of Multiply Imputed Synthetic Data under Multiple Linear Regression Model" (Abhishek Guin, Anindya Roy, and Bimal Sinha) and published as a *Center for Statistical Research and Methodology Research Report Series.* The paper has been submitted to *IJSS (International Journal of Statistical Sciences).*

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

C. 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 FY2022, there was no significant progress on this project.

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

D. 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 FY2022, research under the above project started and is continuing. A paper has been written and submitted to the *Canadian Journal of Statistics.*

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

E. 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 FY2022, research on this topic is under investigation.

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

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 FY2022, staff organized the thirteenth annual *2022 SUMMER AT CENSUS* which brought 11 recognized scholars to the Census Bureau for

1-3 day (virtual) visits covering broad topic themes including: data integration with administrative records, measurement of race & ethnicity, multi-party computing, racial & ethnic diversity, sampling & estimation, spatial demography, statistical machine learning, and statistical methodology. Each scholar engaged in collaborative research with Census Bureau researchers and staff (Center for Optimization & Data Science; Center for Statistical Research & Methodology; Population Division; Center for Behavioral Science Methods; Social, Economic, and Housing Statistics Division; and Center for Economic Studies;) 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, PUBLICATIONS

Arsham, A., Bebu, I., and Mathew, T. (2022). "A Bivariate Regression-Based Cost-Effectiveness Analysis," *Journal of Statistical Theory and Practice, 16*, Article No. 27.

Binder, C., McElroy, T., and Sheng, X. (2022). "Term Structure of Uncertainty: New Evidence from Survey Expectations," *Journal of Money, Credit, and Banking, 54(1),* 39-71.

Chen, B., McElroy, T., and Pang, O. (2022). "Assessing Residual Seasonality in the U.S. National Income and Product Accounts Aggregates," *Journal of Official Statistics, Volume 38, Issue 2*, 399-428.

Ghosh, T., Ghosh, M., Maples, J., and Tang, X. (2022). "Multivariate Global-Local Priors for Small Area Estimation," *STATS, v5*, 673-688. https://www.mdpi.com/2571-905X/5/3/40/htm.

Janicki, R., Raim, A.M., Holan, S.H., and Maples, J. (2022). "Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models with Application to American Community Survey Special Tabulations," *The Annals of Applied Statistics, Volume 16, Issue 1,* 144-168.

Lucagbo, M. and Mathew, T. (2022). "Rectangular Confidence Regions and Prediction Regions in Multivariate Calibration," *Journal of the Indian Society for Probability and Statistics*, *23*, 155–171.

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

Lucagbo, M., Mathew, T., and Young, D. (In Press). "Rectangular Multivariate Normal Prediction Regions for Setting Reference Regions in Laboratory Medicine*," Journal of Biopharmaceutical Statistics*.

McElroy, T. (2022). "Frequency Domain Calculation of Seasonal VARMA Autocovariances," *Journal of Computational and Graphical Statistics, 31(1),* 301-303.

McElroy, T. (In Press). "Casting Vector Time Series: Algorithms for Forecasting, Imputation, and Signal Extraction," *Electronic Journal of Statistics*.

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

McElroy, T. and Politis, D. (2022). "Optimal Linear Interpolation of Multiple Missing Values," *Statistical Inference for Stochastic Processes*, 1-13.

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

McElroy, T. and Roy, A. (2022). "A Review of Seasonal Adjustment Diagnostics," *International Statistical Review, 90(2),* 259-284.

McElroy, T. and Roy, A. (2022). "Model Identification via Total Frobenius Norm of Multivariate Spectra," *Journal of the Royal Statistical Society, Series B, Volume 84*, 473-495.

McElroy, T. and Trimbur, T. (2022). "Variable Targeting and Reduction in Large Vector Autoregressions with Applications to Workforce Indicators," *Journal of Applied Statistics*, 1-23.

Morris, D.S. and Sellers, K.F. (2022). "A Flexible Mixed Model for Clustered Count Data," *Stats: Special Issue on Statistics, Data Analytics, and Inferences for Discrete Data*, *5(1):* 52–69. https://doi.org/10.3390/stats5010004.

Parker, P., Holan, S., and Janicki, R. (2022). "Computationally Efficient Bayesian Unit-level Models for Non-Gaussian Data Under Informative Sampling with Application to Estimation of Health Insurance Coverage," *The Annals of Applied Statistics, Vol 16, No. 2,* 887-904.

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

Raim, A.M., Nichols, E., and Mathew, T. (In Press). "A Statistical Comparison of Call Volume Uniformity Due to Mailing Strategy," *Journal of Official Statistics*.

Rivas, A., Antoun, C., Feuer, S., Mathew, T., Nichols, E., Olmsted-Hawala, E. and Wang, L (2022), "Comparison of Three Navigation Button Designs in Mobile Survey for Older Adults," *Survey Practice, 15(1).*

Trimbur, T. and McElroy, T. (2022). "Modelled Approximations to the Ideal Filter with Application to GDP and Its Components," *The Annals of Applied Statistics, 16(2),* 627-651.

Wang, Z., Ben-David, E., Diao, G., & Slawski, M. (In Press). "Estimation in Exponential Family Regression Based on Linked Data Contaminated by Mismatch Error," *Statistics and Its Interface*.

Wang, Z., Ben-David, E., Diao, G., & Slawski, M. (2022). "Regression with Linked Datasets Subject to Linkage Error," *Wiley Interdisciplinary Reviews: Computational Statistics, 14(4).*

Weems, K.S., Sellers, K.F., and Li, T. (2021). "A Flexible Bivariate Distribution for Count Data Expressing Data Dispersion," *Communications in Statistics - Theory and Methods*, https://doi.org/10.1080/03610926.2021.1999474.

Wright, T. (2021). "From Cauchy-Schwartz to the House of Representatives: Application of Lagrange's Identity," *Mathematics Magazine, Vol 94*, 244-256.

3.2 BOOKS/BOOK CHAPTERS

Nayak, T.K. (2021). "A Review of Rigorous Randomized Response Methods for Protecting Respondent's Privacy and Data Confidentiality," in *Methodology and Applications of Statistics: A Volume in Honor of C.R. Rao on the Occasion of his 100th Birthday*, ed. B.C. Arnold, N. Balakrishnan and C.A. Coelho, New York: Springer, pp. 319-341.

3.3 PROCEEDINGS PAPERS

Sixth International Conference on Establishment Statistics (ICES VI) Proceedings, June 14-17, 2021

- James Livsey, Colt Viehdorfer, and Osbert Pang, "A New Look at Signal Extraction for Manufacturers' Shipments, Inventories, and Orders (M3) Survey."

3.4 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY RESEARCH REPORTS

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

RR (Statistics #2022-01): Osbert Pang, William Bell, and Brian Monsell, "Accommodating Weather Effects in Seasonal Adjustment: A Look into Adding Weather Regressors for Regional Construction Series," February 16, 2022.

RR (Statistics #2022-02): Abhishek Guin, Anindya Roy, and Bimal Sinha, "Bayesian Analysis of Multiply Imputed Synthetic Data Under the Multiple Linear Regression Model," April 4, 2022.

RR (Statistics #2022-03): Eric V. Slud and Darcy Morris, "Methodology and Theory for Design-Based Calibration of Low-Response Household Surveys with Application to the Census Bureau 2019-20 Tracking Survey," June 22, 2022.

RR (Statistics #2022-04): Yehenew G. Kifle, Alain M. Moluh, and Bimal K. Sinha, "Inference About a Common Mean Vector from Several Independent Multinormal Populations with Unequal and Unknown Dispersion Matrices," August 22, 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.

3.5 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY STUDY SERIES
https://www.census.gov/topics/research/stat-research/publications/working-papers/sss.html

SS (Computing #2022-01): Andrew Raim, James Livsey, and Kyle Irimata, "Browsing the 2010 Census SF2 Summary File with R," August 15, 2022.

3.6 OTHER REPORTS

Bell, W.R., McElroy, T.S., Monsell, B.C., Pang, O., McDonald Johnson, K.M., and Chen, B. (2022). "Identifying Seasonality." U.S. Census Bureau.

Gamerman, V., Kolassa, J., Li, J.Z., Natanegara, F., Sellers, K., Talwai, A., Zou, K.H. "ICSA Panel Discusses Partnerships, Collaborations Across Sectors," *Amstat News*, January 2022, p. 16-17. https://magazine.amstat.org/blog/2022/01/01/icsa-panel

Gamerman, V., Kolassa, J., Li, J.Z., Natanegara, F., Sellers, K., Talwai, A., Zou, K.H. "ICSA Panel Discusses Partnerships, Collaborations Across Sectors, Part 2," *Amstat News*, February 2022, p. 5-7. https://magazine.amstat.org/blog/2022/02/01/icsa-panel-part-2

Gamerman, V., Kolassa, J., Li, J.Z., Natanegara, F., Sellers, K., Talwai, A., Zou, K.H. "A Recap of an ICSA 2021 Panel," *ICSA Bulletin*, March 2022, p. 20-23. https://www.icsa.org/wp-content/uploads/2022/03/ICSA_Bulletin_Mar2022.pdf

Wang, Z., Ben-David, E. & Slawski, M. (2022). "Regularization for Shuffled Data Problems via Exponential Family Priors on the Permutation Group," 25 pages (Preprint) https:/arxiv.org/abs/2111.01767.

4. TALKS AND PRESENTATIONS

International Conference on Advances in Interdisciplinary Statistics and Combinatorics (AISC) 2021, University of North Carolina at Greensboro (virtual), October 8, 2021.

- Ryan Janicki, "A Spatial Change of Support Model for Differentially Private Measurements with Application to Estimation of Counts of Persons in AIAN Areas by Detailed Race Groups."
- Darcy Morris, "Missing Data Methods at the Census Bureau with Massive Nonresponse and Observational Data."
- Andrew Raim, "Direct Sampling in Bayesian Hierarchical Models for Privacy Protected Data."
- Eric Slud, "Nonresponse Weight Adjustment in the Census Bureau's Probability and Nonprobability Tracking Surveys."

Poster (virtual), NBER-NSF Time Series Conference, October 16, 2021.

- Tucker McElroy, "Polyspectral Factorization and Prediction for Quadratic Processes."

Invited Seminar (virtual), 2021-2022 Social Justice Speaker Series, Department of Mathematics & Computer Science, Davidson College*,* November 2, 2021.

- Tommy Wright, "Apportionment & Lagrange's Identity."

Seminar (virtual), Washington University, St. Louis, MO, February 23, 2022.

- Tucker McElroy, "On the Construction, Simulation, and Fitting of Second Order Forecastable Processes."

Census Bureau Scientific Advisory Committee Meeting, Washington D.C., March 18, 2022.

- Tucker McElroy (virtual), "Seasonal Adjustment of Time Series During the Pandemic."

Federal Computer Assisted Survey Information Collection (FedCASIC) 2022 Workshops (Virtual), April 5-6, 2022.

- Emanuel Ben-David, "On Improving Self-Response Rates."

Annual Data Science and Analytics Symposium, Bowie State University, Bowie, MD, April 20, 2022.

- Tommy Wright (Keynote Address), "Measuring Our Condition and Behavior; Censuses, Samples, Big Data."

Invited Presentation (virtual) to Statistics Class, University of Texas at Arlington, Arlington, TX, April 25, 2022.

- Mary H. Mulry, "Data Protection at the U.S. Census Bureau."

77th Annual AAPOR Conference, Chicago, IL, May 11-13, 2022.

- Emanuel Ben-David, "Data Analysis after Record Linkage: Sources of Error, Consequences, and Possible Solutions."

Small Area Estimation Conference 2022 (SAE 2022): Small Area Estimation, Surveys and Data Science, University of Maryland, College Park, May 23 – May 27, 2022.

- Gauri S. Datta (Invited), "An Adhoc Pseudo-Bayes Small Area Estimation with Compromise Regression Weights."
- Gauri S. Datta (presented by Adrijo Chakraborty, Invited), "Objective Bayesian Robust Small Area Estimation with Area-Level Data."
- Kyle Irimata, "A Comparative Study of Various MSE Estimators for EBLUPs from Non-normal Fay-Herriot Models."
- Ryan Janicki, "Bayesian Spatial Hierarchical Models for Differentially Private Measurements of Census Tabulations."
- Joseph Kang, "A Tree-assisted Fay-Herriot Models."
- Jerry Maples, "Examining the Assumption of Constant Model Precision in Small Area Share Models."
- Eric Slud, "Design-based Weight Adjustment for Low-response-rate Surveys."

Seasonal Adjustment Practitioners Workshop, Washington D.C., June 8, 2022.

- Tucker McElroy (virtual, tutorial), "Identifying Seasonality."
- Tucker McElroy (virtual, seminar), "Removing Residual Seasonality from GDP."

SPIRAL Colloquium, Summer Program in Research and Learning (SPIRAL), American University, Washington D.C., June 16, 2022.

- Kimberly Sellers, "Regression Analysis for Dispersed Count Data."

Invited Discussion (virtual) to NSF Research Experience for Undergraduates, Southern Methodist University, Dallas, TX. June 21, 2022.

- Mary H. Mulry, "My Career – Trajectory, Decisions, Jobs, and What Was Fun."

4th International Conference on Statistics: Theory and Applications (ICSTA), Prague, Czech Republic, July 29, 2022.

- Kimberly Sellers, Keynote speaker (virtual): "Dispersed Methods for Handling Dispersed Count Data."

Invited Presentation (virtual) to Sixth International Webinar on Recent Advances in Statistical Theory and Applications, Kerala University, June 29 - July 2, 2022.

- Gauri S. Datta, "Bayesian Spatial Models for Estimating Means of Sampled and Non-sampled Areas."

2022 Joint Statistical Meetings, American Statistical Association, Washington, D.C., August 6-11, 2022.

- Kimberly Sellers (Panelist), "Moving Toward Justice, Equity, Diversity, and Inclusion in the Statistical Sciences."
- Tucker McElroy, "A Model Comparison Diagnostic for Differencing Operators Based Upon Multi-Step Ahead Forecast Mean Squared Error Paths."
- James Livsey, "A Nonstationary Time Series Model for Fractional Seasonal Periodicity."
- Anindya Roy and Tucker McElroy, "A Bayesian Marked Point Process Model for Seasonality in Mixed Frequency Data."
- Osbert Pang, "An Empirical Look at Multivariate Signal Extraction for the U.S. Census Bureau's M3 Survey."
- Jerry Maples and Isaac Dompreh, "Using Geographically Weighted Regressions to Assess Variability in Small Area Model Parameters."
- Kyle Irimata, Jerry Maples, Gauri Datta, "Mean Squared Error Estimation for Non-Normal Small Area Models."
- Tommy Wright, "Optimization of a Joint Confidence Region for a Ranking."
- Gauri Datta, "Pseudo-Bayesian Small Area Estimation."
- Adam Hall, "A Spatial Unified Price Index."
- Eric Slud, "Nonresponse Weight Adjustment in the Census Bureau's Probability and Nonprobability Tracking Surveys."
- Martin Slawski, Brady West, and Emanuel Ben-David, "Analysis of Data Combined from Multiple Sources in the Presence of Linkage Error."
- Booline Chen, Kyle Hood, Tucker McElroy, and Thomas Trimbur, "Business Cycle Fluctuations."

Invited Virtual Presentation at the Massive Data Institute (Georgetown University), Washington, D.C., August 25, 2022.

- Yves Thibaudeau, "William E. Winkler: A High-Impact Career in Statistical Computing."

Second Workshop on Time Series Methods for Official Statistics, Paris, France, September 22-23, 2022.

- Tucker McElroy (virtual), "FLIP: A Utility Preserving Privacy Mechanism for Time Series."

Invited Presentation (virtual) International Conference on Role of Advanced Statistical Tools for Sustainable Development, University of Rajasthan, Jaipur, India, September 23-25, 2022.

- Gauri S. Datta, "Pseudo-Bayes Small Area Estimation."
- Bimal Sinha, "Inference About a Common Mean Vector from Several Independent Multinormal Populations with Unequal and Unknown Dispersion Matrices."

University of Florida, Gainesville, FL, September 26, 2022.

- Tucker McElroy (in person), "Parameterization of Reduced Rank and Co-integrated Vector Autoregression."

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

Paul Parker (U.S. Census Bureau Dissertation Fellow), University of California, Santa Cruz, "Computationally Efficient Bayesian Unit-Level Modeling of Non-Gaussian and Complex Survey Data under Informative Sampling," October 12, 2021.

Jun Shao, University of Wisconsin-Madison/U.S. Bureau of the Census, "Analysis with External Information," November 1, 2021.

Tucker McElroy, U.S. Bureau of the Census, "Inference and Prediction for Quadratic Processes," November 9, 2021.

Rebecca Steorts, Duke University/U.S. Bureau of the Census, "Entity Resolution with Societal Impacts in Statistical Survey Methodology, Statistical Science, and Machine Learning," November 15, 2021.

Joseph Kang, U.S. Bureau of the Census, "Causal Inference with Topic Model-Driven Textual Data," November 17, 2021.

Esther Rolf, University of California, Berkeley, "MOSAIKS: A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery, with Application to ACS," November 30, 2021.

Michelle Nixon (U.S. Census Bureau Dissertation Fellow), The Pennsylvania University, "A Latent Class Modeling Approach for Generating Synthetic Data and Making Posterior Inferences from Differentially Private Counts," February 15, 2022.

Wendy Martinez, U.S. Bureau of Labor Statistics, "Ethical Data Science," March 29, 2022.

Emily Shen, MIT Lincoln Laboratory, *SUMMER (Virtually) AT CENSUS*, "Exploring Possible Multi-Party Computation Uses at the Census Bureau (Part 1)," May 19, 2022.

Rafail Ostrovsky, University of California, Los Angeles, *SUMMER (Virtually) AT CENSUS*, "Exploring Possible Multi-Party Computation Uses at the Census Bureau (Part 2)," May 19, 2022.

David Archer, Galois, *SUMMER (Virtually) AT CENSUS*, "Exploring Possible Multi-Party Computation Uses at the Census Bureau (Part 3)," May 19, 2022.

Yaakov Malinovsky, University of Maryland, Baltimore, County, *SUMMER (Virtually) AT CENSUS*, "Unified Approach for Solving Sequential Selection Problems," June 7, 2022.

Yaakov Malinovsky, University of Maryland, Baltimore, County, *SUMMER (Virtually) AT CENSUS*, "Group Testing: Some Results and Open Challenges," June 8, 2022.

Kris Marsh, University of Maryland, College Park, *SUMMER (Virtually) AT CENSUS*, "The Racial Residential Segregation of Black Single Living Alone Households," June 14, 2022.

Corey Sparks, University of Texas at San Antonio, *SUMMER (Virtually) AT CENSUS*, "Beyond Simple Maps – Integrating Space and Time with Bayesian Models," July 11, 2022.

Neda Maghbouleh, University of Toronto, *SUMMER (Virtually) AT CENSUS*, "MENA Americans May Not Be Perceived, Nor Perceive Themselves, to Be White," July 12, 2022.

Esther Rolf, University of California, Berkeley, *SUMMER (Virtually) AT CENSUS*, "MOSAIKS: A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery (and How to Use It)," July 19, 2022.

Amber Crowell, Fresno State University, *SUMMER (Virtually) AT CENSUS*, "New Approaches to Studying Residential Segregation: Considerations of Data, Measurement, and Research Design," July 26, 2022.

Ted Enamorado, Washington University, *SUMMER (Virtually) AT CENSUS*, "Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records," July 27, 2022.

Yan Li, University of Maryland, College Park, *SUMMER (Virtually) AT CENSUS*, "Non-probability Sample Design and Analysis," August 2, 2022.

Ryan Janicki, U.S. Bureau of the Census, "Bayesian Nonparametric Multivariate Spatial Mixture Effects Models with Application to American Community Survey Special Tabulations," August 31, 2022.

Tommy Wright, U.S. Bureau of the Census, "Ranking," September 1, 2022.

Bimal Sinha, University of Maryland Baltimore County/U.S. Bureau of the Census, "About a Common Mean Vector from Several Independent Multinormal Populations with Unequal and Unknown Dispersion Matrices," September 13, 2022.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

Gold Medal Award, U.S. Department of Commerce

- **Chad Russell** – Team Award in Administrative/Technical Support "…for significant improvements in Census's capability to use the commercial cloud to host Title 26 data while dramatically decreasing time to process data by 92%. For the first time, Census achieved the ability to host Title 26 Data in the AWS GovCloud while engineering a high-performance solution to support the processing of Decennial response files. Thus, the team improved Census's ability to provide timely, relevant, and cost-effective data in a secure, elastic, and highly available environment for future Census products.

American Statistical Association Fellow

- **Tucker McElroy (R&M) –** For outstanding contributions to statistical methodology, especially in multivariate time series and seasonal adjustments; for extensive statistical consulting and mentorship; and for service to the profession.

6.2 SIGNIFICANT SERVICE TO PROFESSION

Emanuel Ben-David

- Refereed papers for the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022), *WIREs Computational Statistics, Mathematical Reviews*
- Member, Committee, 2022 W. J. Dixon Award for Excellence in Statistical Consulting
- Co-organizer, Session on "Advances in Unlabeled Sensing and Learning from Unordered Data," Hybrid SIAM Conference on Mathematics of Data Science (MDS22)
- Member, Doctoral Defense Committee, Department of Mathematics and Statistics-UMBC

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*
- Organizer, Invited Session "Some New Results for the Fay-Herriot Model," SAE Meeting at University of Maryland, College Park, May 23-27, 2022
- Organizer, Invited Session "Some Popular Applications in Data Integration," ICSA Applied Statistics Symposium, University of Florida, Gainesville, June 19-22, 2022
- Refereed papers for *Journal of the Royal Statistical Society A, Sankhya, Survey Methodology, Journal of Survey Statistics and Methodology,* and *Journal of the American Statistical Association*

Kyle Irimata

- Refereed papers for *Survey Methodology*

Ryan Janicki

- Refereed papers for *Journal of the Royal Statistical Society, Series A, Journal of Statistical Computation and Simulation, Journal of Survey Statistics and Methodology, Journal of Official Statistics,* and *Journal of Statistical Software*

Patrick Joyce

- Refereed a paper for *Journal of the Royal Statistical Society, Series A*

James Livsey

- Organizer, Time Series Modeling Mixed Frequency Data, Seasonality, and Model Identification Session, *2022 Joint Statistical Meetings*
- Organizer/Chair, Time Series in Federal Statistics Session, *2022 Joint Statistical Meetings*

Jerry Maples

- Refereed book chapter for *Handbook on Poverty Measures*
- Refereed paper for *Journal of the Royal Statistical Society, Series A*
- Member, Expert Review Panel for NASA's Community Response Testing with the X-59 Quiet Supersonic Technology Aircraft, Onsite Meeting at Langley Research Center

Thomas Mathew

- Associate Editor, *Sankhya*
- Associate Editor, *Journal of Multivariate Analysis*
- Associate Editor, *Journal of Occupational and Environmental Hygiene*

Tucker McElroy

- Refereed papers for *Bernoulli, Econometric Theory, Journal of the American Statistical Association, Annals of Applied Statistics,* and *Bundesbank.*
- Member, Zellner Thesis Award Committee, Business and Economics Statistics Section, American Statistical Association
- Representative for Business and Economics Section, Council of Sections, American Statistical Association
- Associate Editor, *Journal of Time Series Analysis*
- Guest Editor, *Journal of Official Statistics*

Darcy Morris

- Associate Editor, *Communications in Statistics*
- Newsletter Editor, Survey Research Methods Section, American Statistical Association
- Reviewed papers for *TEST* and *Statistical Modeling*

Mary Mulry

- Associate Editor, *Journal of Official Statistics*
- Member, Fellows Committee, Survey Research Methods Section, American Statistical Association

Tapan Nayak

- Associate Editor, *Journal of Statistical Theory and Practice*
- Refereed papers for *Survey Methodology* and *Biometrika*

Andrew Raim

- Member, Ph.D. Defense Committee (for 2 students), Department of Mathematics & Statistics, University of Maryland, Baltimore County

Kimberly Sellers

- Associate Editor, *The American Statistician*
- Associate Editor, *Journal of Computational and Graphical Statistics*
- Commissioning Editor, *WIREs Computational Statistics*
- Refereed paper for *Communications in Statistics – Case Studies and Data Analysis*
- Inaugural Chairperson, Justice, Equity, Diversity, and Inclusion (JEDI) Outreach Group, American Statistical Association
- Member, External Nominations and Awards Committee, American Statistical Association

Bimal Sinha

- Associate Editor, *Environmental Modeling and Assessment, Thailand Statistician, Calcutta Statistical Association Bulletin*, and *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 *Annals of Statistics, Journal of the American Statistical Association,* and *Journal of Pharmacokinetics and Pharmacodynamics*

- Member, Small Area Estimation 2022 Conference (SAE2022) Program Committee
- Organizer, Invited Session "Large-magnitude Weight Adjustments", SAE Meeting at University of Maryland, College Park, May 23-27, 2022

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*

Yves Thibaudeau

- Refereed papers for *Journal of Bayesian Analysis, Journal of Official Statistics,* and *Journal of Survey Statistics and Methodology*

Tommy Wright

- Refereed a paper for *American Economic Review*
- Chair, Tukey (1962) and the Subsequent Sixty Years of Data Analysis Perspectives from Government and Social Science Applications Session, *2022 Joint Statistical Meetings*

6.3 PERSONNEL NOTES

- Adam Hall (Ph.D. in statistics, University of Michigan) converted from a postdoc and accepted a permanent research mathematical statistician position in our center.
- Joseph Kang (Ph.D. in statistics, The Pennsylvania State University) transferred to our center as a principal researcher in our Missing Data & Observational Data Modeling Research Group.
- Paul Parker (Ph.D. in statistics, University of Missouri) accepted a research mathematical statistician Schedule A (University of California, Santa Cruz) appointment in our Small Area Estimation Research Group.
- Rebecca Steorts (Ph.D. in statistics, University of Florida) converted from a Schedule A appointment and accepted a permanent principal researcher position in our Record Linkage & Machine Learning Research Group.

APPENDIX A **Center for Statistical Research and Methodology FY 2022**
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		
6550H01	Data Coding/Editing/Imputation		
6550H06	Redistricting Data Program		
6550H08	Data Products Dissemination Preparation/Review/Approval		
6650H01	PES Planning and Project Management		
6650H20	2020 Evaluations-Planning and Project Management		
5350H01	Address Frame Updating Activities		
5350H04	Demographic Frame Updating Activities		
5450H06	Content, Forms Design, & Language		
5450H10	In-Person Enumeration Planning & Support		
5450H20	In-Office Enumeration Planning & Support		
5450H21	Response Data Quality		
5450H23	Response Processing Planning & Support		
5650H02	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. Advances in the Use of Capture-Recapture Methodology in the Estimation of U.S. Census Coverage Error	Mary Mulry	Tom Mule
	4. Supplementing and Supporting Non-Response with Administrative Records	Michael Ikeda	Tom Mule
	5. 2020 Census Privacy Variance	James Livsey	Phil Leclerc
	6. Identifying "Good" Administrative Records for 2020 Census NRFU Curtailment Targeting	Darcy Morris	Tom Mule
	7. Experiment for Effectiveness of Bilingual Training	Andrew Raim	Renee Ellis
	8. Record-Linkage Support for the Decennial Census	Dan Weinberg	Anup Mathur
	9. Coverage Measurement Research	Jerry Maples	Tim Kennel
	10. Empirical Investigation of the Minimum Total Population of a Geographic District to Have Reliable Characteristics of Various Demographic Groups	Kyle Irimata	James Whitehorne
	11. Statistical Modeling to Augment 2020 Disclosure Avoidance System	Andrew Raim	Michael Walsh
	12. Imputation Modeling for Multivariate Categorical Characteristic Data in 2030 Census	Darcy Morris	Tom Mule
	13. Group Quarters Count Expectation Modeling to Ensure Data Quality	Kim Sellers	Andrew Keller
	14. Mobile Questionnaire Assistance: Analysis and Simulation	Andrew Raim	Lisa Moore
	15. Comparison of Probability (RDD) and Nonprobability in a Census Tracking System	Eric Slud	Jennifer Hunter Childs
	16. Experiment on Minimum Height of Text Display on Mobile Survey	Thomas Mathew	Lin Wang
	17. Experiment on Optimal Font Style for Outdoor Viewing of Mobile Survey	Thomas Mathew	Lin Wang
	18. Identifying Optimal Size of Touch Target for Mobile Survey Instruments	Thomas Mathew	Lin Wang
	19. Design Testing for Reporting Zero in a Web Survey	Thomas Mathew	Jonathan Katz
	20. Agreements for Advancing Record Linkage	Rebecca Steorts	Krista Park
6385H70	American Community Survey (ACS)		
	21. Voting Rights Section 203 Model Evaluation and Enhancements Towards 2021 Determinations	Eric Slud	James Whitehorne
	22. Assessing and Enhancing the ACS Experimental Weighting Approach Implemented in 2020 Data Products	Darcy Morris	Mark Asiala

	DEMOGRAPHIC	
TBA	Demographic Statistical Methods Division (DSMD) Special Projects	
	23. 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	
	24. Data Integration	Edward Porter ....Christopher Boniface
7165022	Social, Economic, and Housing Statistics Division Small Area Estimation Projects	
	25. Research for Small Area Income and Poverty Estimates (SAIPE)	Jerry Maples....Wes Basel
	26. Assessing Constant Parameters across Areas in the SAIPE Models	Jerry Maples....Wes Basel
	27. Small Area Health Insurance Estimates (SAHIE)	Ryan Janicki....Wes Basel
	ECONOMIC	
1183X01	General Economic Statistical Support	
1183X90	General Economic Statistical Program Management	
	28. Use of Big Data for Retail Sales Estimates	Darcy Morris....Stephen Kaputa
	29. Seasonal Adjustment Support	Tucker McElroy... Kathleen McDonald-Johnson
	30. Seasonal Adjustment Software Development and Evaluation	James Livsey..... Kathleen McDonald-Johnson
	31. Research on Seasonal Time Series - Modeling & Adjustment Issues	Tucker McElroy... Kathleen McDonald-Johnson
	32. Supporting Documentation & Software for Seasonal Adjustment	James Livsey..... Kathleen McDonald-Johnson
	33. Exploring New Seasonal Adjustment & Signal Extraction Methods	James Livsey....Colt Viehdorfer
	34. Production & Dissemination of Economic Indicators	Adam Hall....Catherine Buffington
	35. Evaluating and Improving the Low Self-Response Score (LRS) with the 2020 Census Data	Emanuel Ben-David..... Joanna Fane Lineback
	PROGRAM DIVISION OVERHEAD	
0331000	*36. Research Computing*	Chad Russell .... Jaya Damineni
	NATIONAL CANCER INSTITUTE	
9401021	*37. Modeling Tobacco Use Outcomes with Data from Tobacco Use Supplement – Current Population Survey*	Isaac Dompreh .... Benmei Liu

APPENDIX B

FY 2022 PROJECT PERFORMANCE MEASUREMENT QUESTIONNAIRE

CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY

Dear

As a sponsor for the FY 2022 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 2022 ***(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 2022 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 2022 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
VACANT

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
- 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
- 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 (Georgetown U.)

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