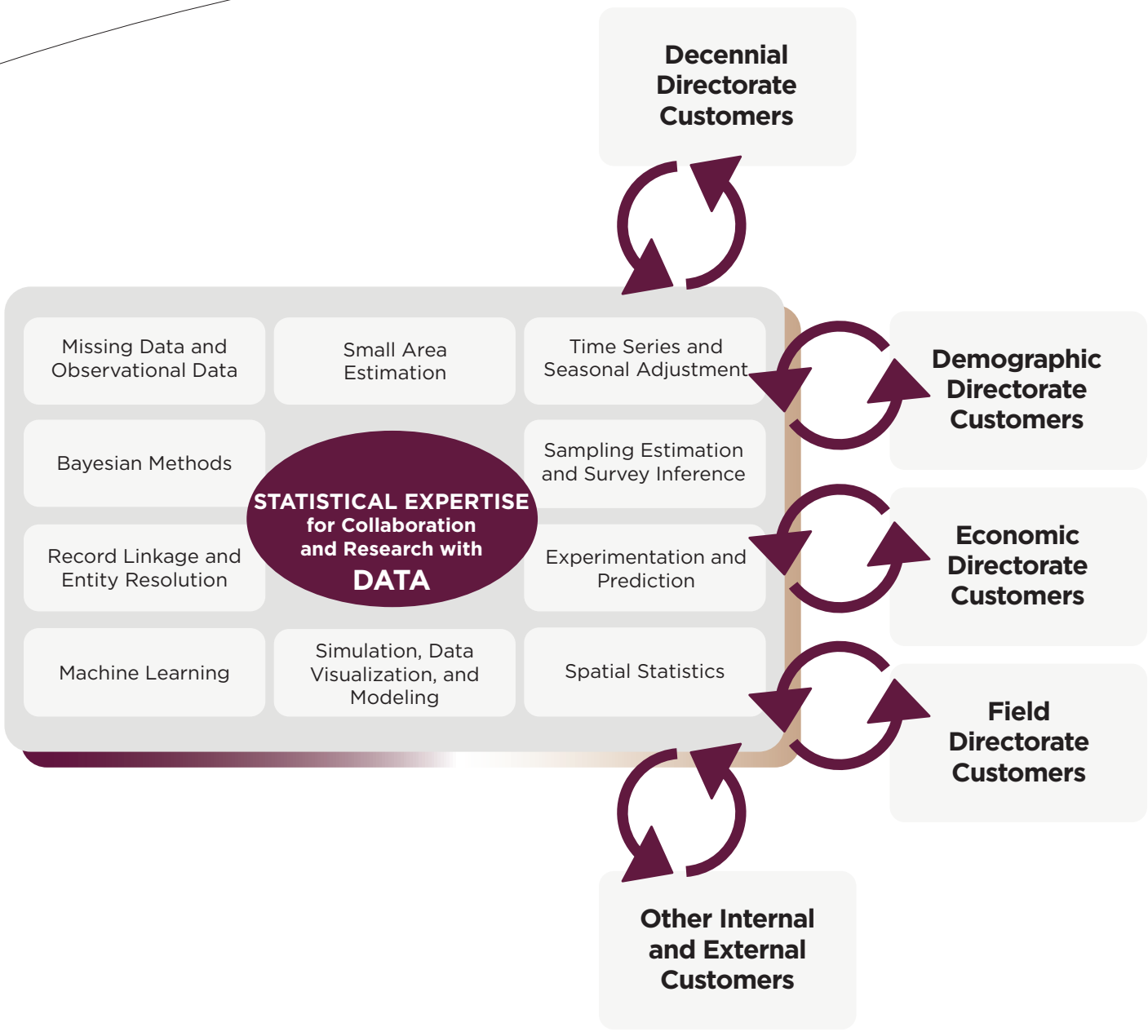
Annual Report of the Center for Statistical Research and Methodology

Research and Methodology Directorate

Fiscal Year 2021



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 2021, 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 2021 follow, and more details are provided within subsequent pages of this report:

- Published with our colleagues in the Decennial Statistical Studies Division a document providing a high-level discussion of the research and methodology underlying the use of administrative records in the 2020 Census 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. Throughout the document, the descriptions of the research and methodology include the rationale behind the resulting decisions [Mulry (CSRM), Mule (DSSD)].

- Conducted research and developed statistical models and appropriate code in STAN and R to meet Section 203, Voting Rights Act requirements to identify those jurisdictions in the United States needing to provide voting materials in languages for those voters who do not speak English well. Work focused on estimating various counts and proportions of various language and minority groups in each jurisdiction as well as incorporating uncertainty measures of sampling variance for the estimators. Many small groups and small American Community Survey sample sizes for many of these groups makes the statistical estimation a challenging and complex task [Franco (CSRM), Slud (CSRM), Others].

- Conducted research and received peer-review consent to publish statistical model-based estimates thereby improving data confidentiality for American Community Survey (ACS) Special Tabulations [Janicki (CSRM), Raim (CSRM), Holan (R&M), Maples (CSRM)].

- Extended and documented previous research on unit-level small area statistical modeling as part of efforts for Small Area Insurance Estimates (SAHIE) sponsored by the Social, Economic, and Housing Statistics Division. A variational Bayes method was implemented to reduce computational time and memory requirements [Janicki (CSRM), Holan (R&M)].

- Worked with Economic Statistical Methods Division to document (in research paper form) the hierarchical Bayesian imputation model methodology for estimating state-level retail sales using establishment-level business register data, state-level economic data and spatial state-level random effects in the Monthly State Retail Sales (MSRS) report. This experimental product was released in FY2020 and is one of the Census Bureau's first efforts to develop a statistical model that blends traditional sample survey data with administrative data and third-party data sources, while producing a new data product measuring our rapidly evolving economy [Morris (CSRM), Hutchinson (EID), Scheleur (EID), Thompson (ESMD)].

- Reported research on seasonal time series-modeling and adjustment issues; (1) continued documentation for the assessment of weather effects on seasonal adjustment, (2) completed review paper on seasonal adjustment diagnostics, (3) developed methodology, code, and empirical results for a maximum entropy framework of extreme value adjustment to handle additive outliers and level shifts in economic time series, and (4) refined empirical results on the use of the EM algorithm for fitting multivariate time series [McElroy (R&M), Livsey (CSRM), Pang (CSRM), Trimbur (CSRM), Bell (R&M)].

- Analyzed and published updated empirical results on reliability of the TopDown Algorithm (TDA) output using the 2020 Census redistricting data production settings version (epsilon = 17.14) of the TDA for all block groups (proxy for districts) in the United States; and also for proxies, used places and minor civil divisions (MCDs), and legislative districts. Empirical results suggest a minimum TOTAL that is between 450 and 499 people in a block group provides reliable characteristics of various demographic groups in a block group based on the TDA. Similarly, a minimum TOTAL that is between 200 and 249 is observed to provide reliable characteristics for places and MCDs. No congressional or state legislative district failed our test for reliability [Wright (CSRM), Irimata (CSRM)].

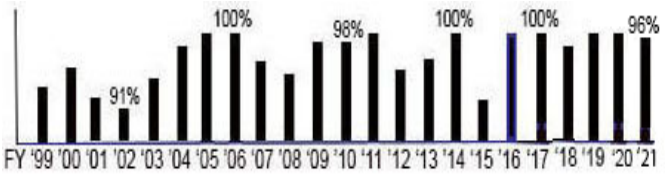
- Analyzed and published updated empirical results on variability of TDA output using the 2020 Census redistricting data production settings version (epsilon = 17.14) of the TDA to the 2010 Census Edited File (2010 CEF) for districts in Rhode Island and for three additional jurisdictions in Mississippi provided by the U.S. Department of Justice. We reported observations on variability of results among 25 independent runs of the TDA, and we reported observations on variability between the results among the 25 runs on the TDA and the published 2010 Census Public Law 94-171 data. Variability with the 2020 Census redistricting data production settings version of the TDA (epsilon =17.14) tends to be less than what we reported earlier with the 2021-04-28 version of the TDA (epsilon = 10.3) [Wright (CSRM), Irimata (CSRM)].

How Did We[1] Do...

For the 23rd year, we received feedback from our sponsors. Near the end of fiscal year 2021, our efforts on 28 of our program (Decennial, Demographic, Economic, Administration, External) sponsored projects/subprojects with substantial activity and progress and sponsor feedback (Appendix A) were measured by use of a Project Performance Measurement Questionnaire (Appendix B). Responses to all 28 questionnaires were obtained with the following results (The graph associated with each measure shows the performance measure over the last 23 fiscal years):

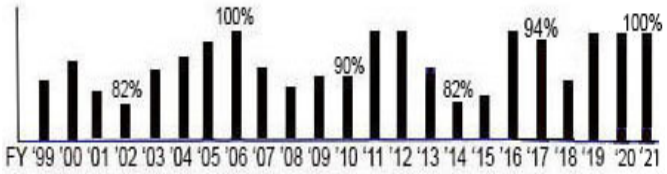
Measure 1. *Overall, Work Met Expectations*

Percent of FY2021 Program Sponsored Projects/Subprojects where sponsors reported that overall work met their expectations (agree or strongly agree) (27 out of 28 responses) ……. 96%



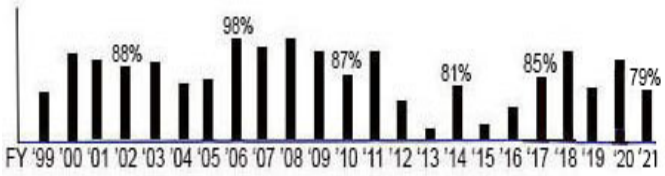
Measure 2. *Established Major Deadlines Met*

Percent of FY2021 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met (17 out of 17 responses) ……..…..…..…... 100%



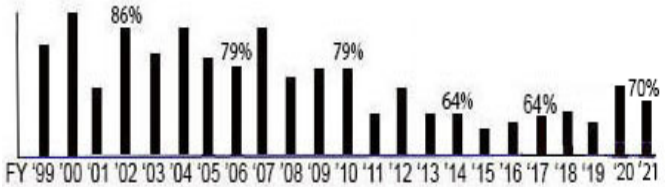
Measure 3a. *At Least One Improved Method, Developed Technique, Solution, or New Insight*

Percent of FY2021 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight (22 out of 28 responses) …… 79%



Measure 3b. *Plans for Implementation*

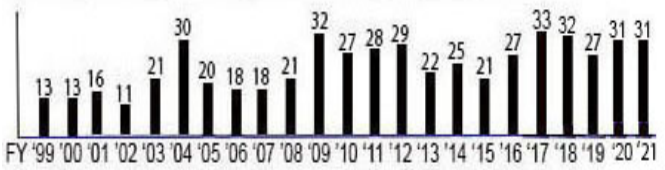
Of these FY2021 Program Sponsored Projects/Subprojects reporting at least one improved method, technique developed, solution, or new insight, the percent with plans for implementation (19 out of 27 responses) ……..…….. 70%



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

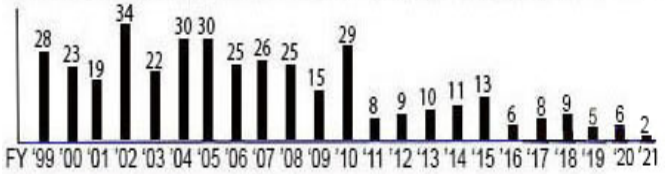
Measure 4. *Journal Articles, Publications*

Number of peer reviewed journal publications documenting research that appeared (18) or were accepted (13) in FY2021 …….. 31



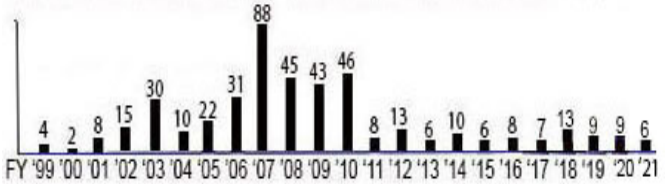
Measure 5. *Proceedings, Publications*

Number of proceedings publications documenting research that appeared in FY2021 ……..……. 2



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

Number of center research reports/studies publications documenting research that appeared in FY2021 ……...…… 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
Decennial Directorate 1
1.1 Project 6550G01 – Data Coding/Editing/Imputation
1.2 Project 6550G06 – Redistricting Data Program
1.3 Project 6550G08 – Data Products Dissemination Prep/Review/Approval
1.4 Project 6650G01 – PES Planning & Project Management
1.5 Project 6650G20 – 2020 Evaluations – Planning & Project Management
1.6 Project 6650G25 – 2030 Planning & Project Management
1.7 Project 6750G01 – Administrative Records Data
1.8 Project 6385G70 – American Community Survey

Demographic Directorate 7
1.9 Project TBA – Demographic Statistical Methods Division Special Projects
1.10 Project 0906/1444X00 – Demographic Surveys Division (DSD) Special Projects
1.11 Project 7165021 – Social, Economic, & Housing Statistics Division Small Area Estimation Projects

Economic Directorate 9
1.12 Project 1183X01 – General Economic Statistical Support
1.13 Project 1183X90 – General Economic Statistical Program Management
1.14 Project 2120G90 – Economic Census Program Management

Census Bureau 12
1.15 Project 0331000 – Program Division Overhead
1.16 Project TBA – National Cancer Institute Special Projects

2. RESEARCH 15
2.1 Project 0331000 – General Research and Support
Missing Data & Observational Data Modeling
Record Linkage & Machine Learning
Sampling Estimation & Survey Inference
Small Area Estimation
Time Series & Seasonal Adjustment
Experimentation, Prediction, & Modeling
Simulation, Data Science, & Visualization
SUMMER AT CENSUS
Research Support and Assistance

3. PUBLICATIONS 27
3.1 Journal Articles, Publications
3.2 Books/Book Chapters
3.3 Proceedings Papers
3.4 Center for Statistical Research & Methodology Research Reports
3.5 Center for Statistical Research & Methodology Study Series
3.6 Other Reports

4. TALKS AND PRESENTATIONS 30

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES 32

6. PERSONNEL ITEMS 33
6.1 Honors/Awards/Special Recognition
6.2 Significant Service to Profession
6.3 Personnel Notes

APPENDIX A
APPENDIX B

1. COLLABORATION

1.1 DATA CODING/EDITING/IMPUTATION (Decennial Project 6550G01)

1.2 REDISTRICTING DATA PROGRAM (Decennial Project 6550G06)

1.3 DATA PRODUCTS DISSEMINATION PREPARATION/REVIEW/APPROVAL (Decennial Project 6550G08)

1.4 PES PLANNING & PROJECT MANAGEMENT (Decennial Project 6650G01)

1.5 2020 EVALUATIONS – PLANNING & PROJECT MANAGEMENT (Decennial Project 6650G20)

1.6 2030 PLANNING & PROJECT MANAGEMENT (Decennial Project 6650G25)

1.7 ADMINISTRATIVE RECORDS DATA (Decennial Project 6750G01)

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. A document is desired to provide 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 FY2021, staff completed a collaboration with staff in the Decennial Statistical Studies Division on the preparation of the document entitled "Overview of Administrative Records Modeling in the 2020 Census." The 2020 U.S. Census is the first U.S. Census to use administrative records (ARs) to enumerate some households. The document provided 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 document was posted on census.gov in April 2020 close to the time of the release of the 2020 Census counts for states that would be used in apportionment of the seats in the U.S. Congress. The document provided an explanation of the AR enumeration methodology that was accessible to a wide audience. The final version of the document was entered in the 2020 Census Program Memorandum Series with the number 2021.10 and posted on census.gov.

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 FY2021, staff began collaborating with staff in the Decennial Statistical Studies Division (DSSD) on implementing the study that compares 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.

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 FY2021, staff continued to document results of implementations of an outlier detection methodology on the housing unit (HU) enumeration universe on the 2020 Decennial Enumeration Master Address file. Staff documented the results in five draft memoranda (which are in the process of being turned into Decennial Statistical Studies Division (DSSD memoranda) and three outlines. The first memorandum covers the methodological history and basic results. There is some tendency for high ratio object scores (the score from the outlier detection methodology) to be associated with indications of questionable unit status. The second memorandum contains an overall comparison of results from the outlier detection methodology to results from the 2020 production administrative records (AR) modeling methodology. High ratio object scores are associated with AR modeling deletes and (to a lesser extent) with AR modeling vacant status. The third memorandum concentrates on results for AR modeling deletes and the fourth memorandum looks at AR modeling deletes by ratio object score category (the categories are based on percentiles of the ratio object score). AR modeling deletes tend to have values of individual outlier detection modeling variables suggesting questionable unit status, although this is less true for AR modeling deletes with relatively low ratio object scores. The fifth memorandum looks at AR modeling vacants and deletes by whether they were former AR modeling vacant/delete overlap cases. Former AR modeling vacant/delete overlap cases converted to AR modeling deletes are like regular AR modeling deletes in their ratio object score distribution, although there are differences in the distribution of some outlier detection modeling variables. Former overlap cases converted to AR modeling vacants are in between regular AR modeling vacant cases and former overlap cases converted to AR modeling deletes in their ratio object score distribution and often in their distribution of individual outlier detection modeling variables. The first outline compares AR modeling closeout vacants and deletes to AR modeling vacants and deletes. Compared to AR modeling deletes, AR modeling closeout deletes tended to have somewhat lower ratio object scores and a tendency for individual outlier detection modeling variables to be shifted away from values suggesting questionable unit status. In contrast, AR modeling closeout vacants tended to be like AR modeling vacants in the distributions of both ratio object scores and individual outlier detection modeling variables. The second outline looks at results for Puerto Rico. Because AR modeling did not assign unit status in Puerto Rico, a pseudo-status was assigned using the indicator variables from AR modeling for meeting the occupied, vacant, and delete removal thresholds. 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 outline compares outlier detection results to 2020 Census Unedited File (CUF) results. There appears to be a tendency for CUF units with final status of delete (excluding units with status assigned by AR modeling) to have higher ratio object scores and for their distributions of individual outlier detection modeling variables to be shifted towards values suggesting questionable unit status. Staff wrote two additional draft memoranda. The first memorandum compares tract-level results from the outlier detection methodology to calculations by staff of the Demographic Statistical Studies Division (DSSD) of tract-level differences between Tax Year 2019 and Tax Year 2018 Internal Revenue Service (IRS) form 1040 response rates. There seems to be a tendency for higher tract ratio object score third quartiles to be associated with higher percentiles of tract-level differences between Tax Year 2019 and Tax Year 2018 IRS form 1040 response rates. The second memorandum documents the analysis of files of off-campus residents obtained from various colleges for their potential for use in unduplication between local (near-campus) and alternate (possibly parental) addresses. The files did not appear to be particularly useful for this purpose. Finally, staff fit four multinomial logistic regression models on 2010 AR HU data for non-response follow-up (NRFU) units (with 2010 CUF) 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 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. The number of "selected" units was chosen to match (separately for units with predicted HU size of 0 and predicted HU size greater than 0) the number of units with status assigned by AR modeling. The first model was based on IRS 1040 household size with separate sub-models for units with different values of the relevant IRS 1040 household size. This model is referred to as the "initial" model. The second model is the "combined" model which is like the "initial" model except that separate sub-models are not used for different values of the relevant IRS 1040 household size. The third model has separate sub-models

for different values of AR household composition. The fourth model has separate sub-models for different values of AR household size. The "initial" model generally does better both in agreement to the CUF population total for selected units (both all selected units and selected units with status not assigned by AR modeling) and in accurate prediction of the CUF HU size (for selected units where HU status is not assigned by AR modeling). Staff documented these results in four outlines, one for each of the models. The outline for the latter three models also compared results from their given model to the "initial" 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 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 FY2021, data were retrieved from the Census Bureau Planning Database (PDB) and Population Estimates Program (POP). Two way interactions at the tract level are accessible from the PDB while high level interactions are accessible at the county level from POP. Relevant portions of each database were used as raking variables against a trivial array to create a realistic starting population. This allowed for realistic multiway interaction while still being able to control the inputs in a simulation study.

A factorial simulation study was designed to successively step from an empty workload up to a full detailed workload. Code to implement the planned study is in development. Finalized study and code will include complete workflow starting with Log-linear model inputs and survey calibration of realistic multi-way array and ending at cell-by-cell variance estimates resulting from L1 regression modeling.

Initial runs of factorial simulation were completed. These initial runs were then studied in-depth. This investigation considered the slack variables of the L1 regression and how they impact estimates. This information was used to tune future replicates of the simulation study.

Slack variables that result from least absolute regression fits were further explored. Most focus was concentrated on their relationship to the number of marginals each detailed cell was a part of. This surmounts to understanding the max/min workload and corresponding privacy budget schedule that would allow the best possible way to economize resources when calculating the variances.

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 FY2021, staff worked with Decennial Statistical Studies Division and Center for Economic Studies colleagues on documenting 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.

Staff: Darcy Steeg Morris (x33989), Yves Thibaudeau

E. Experiment for Effectiveness of Bilingual Training
Description: Training materials will be 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 a statistical experiment to be 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 FY2021, Center for Behavioral Science Methods (CBSM) staff discovered that, during census operations, the updated training module had been assigned to enumerators differently than prescribed in sample size planning. Center for Statistical Research and

Methodology proposed a revised model for analysis while CBSM staff specified the content and format of the data to be used for the analysis. The main experimental question (effectiveness of the training) and a number of exploratory analyses of interest will be carried out once the data become available.

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 FY2021, there was no progress on this project. The project will resume and generate a final report in the first half of FY2022.

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 FY2021, staff continued the discovery process for record-linkage solutions for Census 2030. Staff configured the vendor record-linkage application "Senzing" which was the object of production testing. Staff negotiated upgrades to Senzing from the vendor and requested upgrades to the AWS research cloud from Administrative & Customer Services Division (ACSD). ACSD and cloud engineering provided a large capacity relational database supported by 1.5 TB of low-latency NVMe storage. Senzing for entity resolution is a change of paradigm relative to traditional record-linkage software. To resolve entities many record-linkage software block the records that are candidates for entity resolution by geography, date of birth, etc. before making comparisons between records in a same block. Senzing eschews blocking. Senzing accumulates records in a relational database and subsequently queries the database for entities resembling the most recently processed records. So, the latency of the relational database being queried is crucial to the performance of Senzing. In that context, new standards must be developed to compare Senzing to record-linkage software involving blocking. Center for Statistical Research and Methodology/Center for Optimization and Data Science staff jointly taught a class in record matching and record linkage to about 100 students in the "Data Science" curriculum at the Census Bureau. Staff prepared material and demonstrational python programs for record linkage.

Work was undertaken on linking the unPIKed records on the 2020 CUF to the Enhanced Reference File (ERF), which is a composite consisting of over one billion administrative and commercial records, using BigMatch. Preliminary results show that between 1 and 2% of unPIKed records can be found on the ERF. This work supports the 2020 administrative record census exercise and provides groundwork for conducting administrative record censuses in the future.

Eventually the linkage of the 2020 CUF to the ERF was completed with about 1.8% of the unPIKed records linked. Most of the CUF records did not contain sufficient name or DOB information to perform a successful linkage.

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 FY2021, the Post Enumeration Survey (PES) was in the deployment stage and staff is waiting for the data to become available.

Staff: Jerry Maples (x32873), Ryan Janicki

I. Assessing Variability of Data Treated by TopDown Algorithm for Redistricting

Description: Data from the most recent decennial censuses are used by the U.S. Department of Justice (DOJ) to clear some new proposed redistricting plans of U.S. House of Representatives congressional districts, as well as for some new state level districts. The objective of this study is to assess the variability of data results from the application of a disclosure avoidance randomization algorithm to 2010 Census Edited File Data (CEF) for Rhode Island.

Highlights: During FY2021, staff analyzed and published updated empirical results on variability of TopDown Algorithm (TDA) output using the 2020 Census redistricting data production settings version (epsilon = 17.14) of the TDA applied to the 2010 Census Edited File (CEF) for districts in Rhode Island and for three additional jurisdictions in Mississippi provided by the U.S. Department of Justice. Staff reported observations on variability of results among 25 independent runs of the TDA, and staff reported observations on variability between the results among the 25 runs of the TDA and the published (based on swapping) 2010 Census Public Law 94-171 redistricting data. Variability with the 2020 Census redistricting data production settings version of the TDA (epsilon = 17.14) tends to be less than what was reported earlier with the 2021-04-28 version of the TDA (epsilon = 10.3). (See Wright and Irimata, May 2021 and August 2021.)

Staff: Tommy Wright (x31702), Kyle Irimata

J. 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 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 FY2021, staff analyzed and published updated empirical results on reliability of the TopDown Algorithm (TDA) output using the 2020 Census redistricting data production settings version (epsilon = 17.14) of the TDA for all block groups (proxies for districts) in the United States; and also for proxies, used places and minor civil divisions (MCDs) and legislative districts. Empirical results suggest a minimum TOTAL that is between 450 and 499 people in a block group provides reliable characteristics of various demographic groups in a block group based on the TDA. Similarly, a minimum TOTAL that is between 200 and 249 people is observed to provide reliable characteristics for places and MCDs. No congressional or state legislative district failed our test for reliability. For details on what is meant by reliability, see the report from which staff shares the following brief quote: *"for any block group with a TOTAL count between 450 and 499 people, and for MCDs and places between 200 and 249, the difference between the TDA ratio of the largest demographic group (LDG) and the corresponding SWA ratio for the LDG is less than or equal to 5 percentage points at least 95% of the time. No Congressional or state legislative district fails this test; that is for these districts, the 5 percentage point criterion holds 100% of the time."* (See Wright and Irimata, May 2021 and August 2021.)

Staff: Tommy Wright (x31702), Kyle Irimata

K. Comparing Swapping Records Results with Differential Privacy

Description: Under this project, staff will compare two algorithms that added noise to the Census Edited File of the 2010 Census. The first algorithm used a Swapping technique (SWA) via Disclosure applications. The second algorithm applied noise using the Differential Privacy technique (TDA) on 4 different occasions: 2019-10-31, 2020-05-27, 2020-09-17 & 2020-11-16. We will compare data for the entire population and data for those over 18 years of age, separated by 63 race categories.

Highlights: During FY2021, staff considered two data files. The SWA data file consists of a record for each block in a state and contains the count for each of the 63 race categories for that block. The TDA data file consists of a record for each person in a state with a field containing a code that represents with which race category that person identifies. Work began by merging data from all Census blocks in the Maryland data sets from the 2021 Census SWA data and the 2020-11-16 TDA data. Staff looked at discrepancies between the 2 data sets for each block in the District of Columbia. Eventually, a library of tables has been created for selected states. These tables include an SWA record for each block in that state and contains the count for each of the 63 race categories for that block. These tables also include a TDA record for each person in a state with a field containing a code that represents with which race category that person identifies.

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

L. 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 FY2021, staff investigated Bayesian hierarchical models for producing more accurate estimates than those produced directly after application of differential privacy. Staff conducted simulation studies to show that the continuous Gaussian distribution is a suitable approximation to Laplace or discrete Gaussian noise mechanisms in such models. Studies utilized the Bayesian modeling platform STAN as well as Gibbs sampling with direct sampling steps. A manuscript detailing this work is in preparation. Staff also investigated change of support methods for producing model-based estimates in target geographies where direct estimates are not released.

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

1.8 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 6385G70)

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 FY2021, staff devised a regression model to estimate veteran population by county and met with Veterans Affairs to discuss implementation. Staff discussed the comparability of multi-year estimates with a researcher from the U.S. Government Accountability Office.

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

B. Assessing Uncertainty in ACS Ranking Tables
Description: This project presents results from applying statistical methods which provide statements of how good the rankings are in the ACS Ranking Tables [See The Ranking Project: Methodology Development and Evaluation, Research Section under 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 FY2021, staff developed STAN code for two parametrizations of a Multinomial Logit Normal (MLN model), and for temporal versions of both MLN models. Staff also developed STAN code to make proportion predictions for small areas with no ACS data. Staff also developed R code to fit the non-temporal MLN models from a frequentist approach using maximum likelihood. Staff tested both codes for convergence using data from several LMGs and via simulations. Staff also did sensitivity analysis on the choice of priors of Bayesian MLN code. Staff compared Frequentist and Bayesian code to check that they give consistent results.

Staff devised potential models for modeling voting age persons by LMG and Geography. Staff considered alternatives for incorporating uncertainty of sampling variance of estimators of population totals into final modeling strategy.

Staff worked to refine and update the SAS preprocessing code, and to improve the code and its documentation.

Staff worked with Center for Optimization and Data Science (CODS) staff who are responsible for the

computation. Staff met regularly with CODS staff to coordinate work and to set project deadlines.

Staff fit two versions of a Dirichlet Multinomial model and compared the predictions to those of the MLN2 model. Staff performed covariate selection, using covariates drawn from ACS, computed at higher levels of aggregations than that of the response. Staff examined model diagnostics. Staff prepared the STAN and R Code for production. Staff made several test runs of the code. Staff also oversaw the work of programmers from CODS.

The project culminated in quarter 4 with the delivery of the estimates needed to support the 2021 Determinations of Mandatory Ballot Assistance by the Acting Director of the Census Bureau under Voting Rights Act Section 203(b). These estimates consisted of model-based estimates of counts within Language Minority Groups (LMGs) of Citizens, Limited English Proficiency (LEP) Citizens, and Illiterate LEP Citizens, and of certain ratios of these counts, within political subdivisions (County/MCD units, American Indian Areas, or Alaska Native Regional Corporations). The models were developed by Center for Statistical Research and Methodology (CSRM) staff (Franco and Slud), implemented in R and STAN batch computer code by them with the assistance of CODS staff (Kang and Al Attar), and tested by these staff together with Decennial Statistical Studies Division (DSSD) staff (under Asiala). Geocoding and other data preprocessing steps were aided by CODS and DSSD staff. The models and numerical analyses and tests developed represented original methodological research by CSRM staff, and additional model analyses were assisted by CSRM staff (Lu and Hall). The estimates delivered included estimated counts and ratios as well as estimated variances of these primary quantities. CSRM staff (Franco) delivered briefings to the Associate Director for Decennial and Acting Census Bureau Director on the overall methodology used in producing these data. Work on technical documentation of the methodology and software documentation continues into FY2022.

Staff: Eric Slud (x34991), Carolina Franco, Adam Hall, Xiaoyun Lu, Mark Asiala (DSSD), Joseph Kang (CODS), Patrick Campanello (CODS), Ahmed Abdulkarim Al Attar (CODS), Tommy Wright

D. Model-based Estimates to Improve Data Confidentiality for ACS Special Tabulations

Description: ACS special tabulations are custom data releases requested by external customers. The released tables, which are often based on small sample sizes, raise concerns with data privacy and confidentiality. This project is to create model based estimates of the special tabs.

Highlights: During FY2021, staff received referees' comments and worked to address all of the raised issues on previously reported work on two different Bayesian methods to attempt to estimate the number of spatial mixture components that are needed to fit some data. Staff submitted a revised manuscript which was accepted for publication in the *Annals of Applied Statistics*.

Staff: Jerry Maples (x32873), Andrew Raim, Ryan Janicki, Scott Holan (R&M), Tommy Wright, John Eltinge (R&M), John Abowd (R&M)

1.9 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 FY2021, the simulation code was further developed and polished to explore biases in SDR variance-estimates under complex survey structures well beyond those investigated by previous authors. Due to some surprising results encountered, showing biases greater in magnitude than those seen in previous published simulations, staff did extensive benchmark testing to confirm the correctness of the previous published results and identify some of the complex-survey features that result in greater biases of SDR variance estimation in small domains. Further simulation results are being collected in early FY2022 and will be written up in internal technical reports and journal papers.

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

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

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

Highlights: During FY 2021, staff wrote data integration and updated Record Linkage software. Software also included updating the social network of links project, which helps reduce "false matches" by Record Linkage software. The software developed by the staff supported the release of 2020 Decennial Microdata with preparation and validation of test data. The results were presented in "U.S. Census Bureau's Ex Post Confidentiality Analysis of the 2010 Census Data Publications."

Staff: Ned Porter (x31798)

1.11 SOCIAL, ECONOMIC, AND HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165021)

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 FY2021, staff created a spinoff project (see Project B in this section) to create an evaluation framework to explore how the parameters of the Fay-Harriot in the production SAIPE models (state and county) vary over features in the data, e.g., geography or population sizes. Staff also started development of a small area model for correlated shares to estimate the number of poor and non-poor children in school districts. Staff is pursuing a Bayesian approach which uses the fact that normalized Gamma random variables with the same scale factor have a Dirichlet distribution. Also, since the sum of two gamma distributed random variables with the same scale has a gamma distribution (with the size terms combined), the common component between the two set of shares can be specified. Staff is still deriving the conditions for the prior distributions to guarantee a proper posterior distribution of the parameters.

Staff: Jerry Maples (x32873), Carolina Franco, 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 FY2021, staff developed a framework to explore how the parameters of the Fay-Harriot vary over subspaces defined by a set of covariates, e.g., geography. The goal is to create an evaluation tool to check the SAIPE production models for lack of fit. Staff extended the methodology used in GWR (graphically weighted regressions) to the small area estimation. Staff have designed a method to select the bandwidth, parameter which controls the smoothing weights, which minimizes the mean squared error of prediction rather than the traditional cross validation measure on the regression fit.

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.

<u>Development of unit-level small area modeling strategies under informative sampling designs.</u>

Highlights: During FY2021, staff continued work on unit-level small area modeling using pseudolikelihoods. Staff extended previous work to accommodate Binomial and Multinomial response data. A Polya-Gamma data augmentation technique was used to ensure all full conditional distributions belong to standard parametric

families. A variational Bayes method was implemented to reduce computational time and memory requirements. A paper documenting this work was accepted for publication in *Annals of Applied Statistics*.

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

1.12 GENERAL ECONOMIC STATISTICAL SUPPORT (Economic Project 1183X01)

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

1.14 ECONOMIC CENSUS PROGRAM MANGEMENT (Economic Project 2120G90)

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 FY2021, staff worked with Economic Statistical Methods Division (ESMD) staff on documenting – in research paper form – the hierarchical Bayesian imputation model methodology for estimating state-level retail sales using establishment-level business register data, state-level economic data and spatial state-level random effects in the Monthly State Retail Sales (MSRS) report. This experimental product was released in FY2020 and represents a step toward providing more timely, granular, and relevant data products that meet data user needs, while minimizing the burden on respondents. The MSRS is one of the Census Bureau's first efforts to develop a model that blends traditional survey data with administrative data and third-party data sources, while producing a new data product measuring our rapidly evolving economy.

Staff: Darcy Steeg Morris (x33989), Rebecca Hutchinson (EID), Jenny Thompson (ESMD), Stephen Kaputa (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 FY2021, staff provided seasonal adjustment and software support for users within and outside the Census Bureau and added new examples to a document that provides recommendations on assessing residual seasonality in economic time series. Staff gave specific seasonal adjustment support to National Accounts of Peru, BGC Liquidez, the Federal Reserve Bank of Kansas, the European Commission, Nifty, Woodstat, NYC Office of Comptroller, NYC Office of Management and Budget, Bureau of Labor Statistics, and the University of Basel. Staff developed a GitHub site with repositories for Ecce Signum, and managed inquiries regarding the software.

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 CODS and ESMD.

Highlights: During FY2021, work was conducted on both the current seasonal adjustment software X-13ARIMA-SEATS (X-13) as well as planning for future variants.

The current codebase was updated and bugs fixed. A new build of X-13 was released internally (Build 58) for testing. Additionally, through collaboration with Census Bureau IT we have begun to monitor user downloads.

Internal testing of Build 58 of X-13 was conducted. No issues were found with building permits time series or Survey of Construction.

X-13 functionality was updated such that usage of 0.0 to denote end-of-series for sequence outliers (AOS/LSS).

A prototype of future seasonal adjustment software continued to be developed. X-13 code was ported to Python and a GUI developed. The X-13 development staff demonstrated a prototype python program. This initial version outlined the team's GUI ideas about including specs within runs of X-13. Feedback was taken from all team members and incorporated.

Sigex development was continued. Specific additions to the software include relevant code to implement fractions seasonal periodicities.

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 FY2021, staff made progress on several research projects: (a) continued documentation for the assessment of weather effects on seasonal adjustment; (b) completed review paper on seasonal adjustment diagnostics; (c) completed empirical assessment of benchmarking optimization methodology for indirect seasonal adjustment of quarterly time series; (d) completed assessment of residual seasonality in GDP using seasonality diagnostics; (e) completed empirical results on the use of the EM algorithm for fitting multivariate time series; (f) extended outlier framework to simultaneously allow for missing values and generic types of outliers; (g) developed methodology, code, and empirical results for a maximum entropy framework of extreme value adjustment, to handle additive outliers and level shifts in economic time series.

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

E. Supporting Documentation and Software for Seasonal Adjustment

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

Highlights: During FY2021, staff revised the manual and quick reference guide to coincide with new release of X-13ARIMA-SEATS (build 58). This included adding a description of the 0.0 convention for outlier sequence regressors and updating bibliography reference. Table names were updated to ensure consistency across appendices and quick reference guide.

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

F. Redesign of Economic Sample Surveys (Stratification)

Description: Following the recommendations of a National Academy of Sciences panel, work was begun to redesign the economic sample surveys into a common sampling and estimation system. This process seeks to take several economic sample surveys and reformulate the surveys as part of a singular sampling and estimation operation. Separate research teams were formed with different research tasks. The work of this project involves

a single research team focused upon the construction of a stratification methodology for the common economic survey execution.

Follow-up work is expected to focus on the practical application of stratification when dealing with multiple responses of interest and multiple variables to apply stratification techniques in order to minimize coefficients of variation. Specifically, the focus is on using two (or more) available administrative sources to relate to a single response or multiple responses. This is to be considered alongside the multitude of NAICS classifications and appropriateness of use.

Highlights: During FY2021, the activities of the Stratification Research Team consisted of regular meetings through the late summer of FY2021, in which we discussed the use of ACES, ARTS and AWTS data in analyzing the variances of survey outcome-variables under various alternative stratifications. We considered primarily stratification in terms of MOS variables and NAICS groupings, and we attempted to formulate optimization problems whose solutions could lead to optima stratifications, based on models informed by data analysis. We produced data-analytic exhibits and writeups on possible objective functions for optimization and used these in briefings for the Steering Committee for the Economic Survey Redesign. However, decisions taken by that Steering Committee have changed the nature of the possible stratification, with the goal of producing estimates from a unified economic survey down to the state level in NAICS-based strata. This small research team discontinued its meetings and research efforts at the end of summer FY21, pending further instructions from the ECON Steering Committee.

Staff: Eric Slud (x34991), Patrick Joyce, Lucas Streng (ESMD), Justin Smith (ESMD)

G. 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 FY2021, staff built upon FY2020 revision history and our custom mimic production code. For revisions, we investigated size based on filter length. Code was written to convert SEATS canonical decomposition to Ecce Signum form. This allowed us to show equivalence in both methods of extracting seasonal and trend signals. The implications of this result are not only important to the current M3 project but any additional multivariate signal extraction research done at the Census Bureau.

The lead and lag structure of M3 series entering the pandemic period was investigated. This entailed observing the exact time period each M3 series began to demonstrate a turning point; a key feature when looking for improvement when jointly modeling series.

Our custom code converting SEATS canonical decomposition to Ecce Signum form was further optimized. This included automated reading of X-13ARIMA-SEATS unified diagnostics file (UDG). This work provided initial values for all multivariate modeling efforts.

The inductive process of fine tuning our multivariate structure component model was started. This involves choosing proper vector SARIMA structure and differencing operators for trend and seasonal components and then looking at model diagnostics to optimize.

A proceedings paper was submitted to the ICES conference. All results for SEATS and multivariate decompositions were documented and upcoming goals outlined.

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

H. 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 FY2021*,* staff performed supervised and unsupervised record linkage to link Google API data that contains detailed information about businesses to Business Register (BR). The matches were analyzed and compared for assessing quality, estimating match rate, and false and missed match rates. We used machine learning techniques to build a classification model that predicts the NAICS for each business in Google API data. We used the linked data for training and validating this model. We presented the initial findings to an internal machine learning group and more comprehensive results at the Sixth International Conference of Establishment Statistics (ICES VI) on June 16, 2021.

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

I. Production and Dissemination of Economic Indicators

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

Highlights: During FY2021, staff produced a draft paper proposing new methods to measure geographic cost of living based on retail barcode scanner data. The paper builds on Redding and Weinstein's "CES Unified Price Index" to compare the cost of groceries between counties, while accounting for the impact of differences in consumer preferences and the good varieties available in each county. Staff presented the results in this paper at a Center for Statistical Research and Methodology seminar in October 2020. Staff also reviewed evidence that federal economic indicator data are leaking to financial markets prior to release, briefed Economic Directorate leadership on its contents, and made recommendations to preserve public confidence. In response to this review, staff began sitting in on the clearance prerelease meetings for the advance retail sales estimates, to better understand the release procedures behind these numbers.

Staff also studied literature about the design of the Annual Retail Trade Survey, Monthly Retail Trade Survey, Service Annual Survey, and the experimental Monthly State Retail Sales estimates. Read papers concerning a new method of optimal sample allocation between strata that does not require rounding.

Staff: Adam Hall (x32936)

1.15 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 FY2021, staff continued to maintain and support the Integrated Research Environment (IRE)

and the Cloud Research Environment (CRE) prototype. CSVD reconfigured the IRE test cluster to support the testing needed to formulate an upgrade strategy for the RedHat operating system and PBS Professional. We tested the PBS upgrade process from the currently used version to version 2020.1.4, the latest version that supports all three RedHat operating systems in use at the Census Bureau. The IRE was used to host the second cohort of the Data Science Training Program. On the CRE, we provisioned several new projects, including projects supporting the use of formal privacy in the Decennial Census, a project involving modelling approaches to support Section 203(b) determinations of the Voting Rights Act, and a project to test Senzing, a commercial software package for entity resolution. We performed testing of rhdf5 and rhdf5lib, two R packages related to HDF5, a candidate file format for storing "noisy" 2020 DAS measurements. We maintained software not centrally maintained by the IT Directorate, and we upgraded NoMachine Terminal Server, SAS, and Matlab.

Staff: Chad Russell (x33215)

C. Experiments to Improve Web Surveys: Design and Analysis

Description: With the growing prevalence of using online surveys as a mode of data collection, experiments are being planned by the Center for Behavioral Science Methods (CBSM) with the goal of reducing respondent burden and improving sample survey response rates; in particular, for reducing item nonresponse rates. Most respondents are choosing to respond to surveys online, many of which are using their smartphones to do so. Consulting help was provided to CBSM staff on the following two experiments:

(i) Accessing online surveys often requires respondents to read and transpose a URL and alphanumeric ID code from a paper invite they received in the mail into their smartphone. This process introduces undue burden on the respondent and increases the possibility of human error during manual data entry. Including a quick response (QR) code on the paper invite to an online survey could improve this process. One experiment that is being planned is on the development a method for using QR codes to access surveys, and to test their usability compared to manual data entry.

(ii) A second experiment that is being planned is on the placement and design of navigation buttons, and their effects on key outcomes, such as breakoff rates and survey completion times. A 2^6 factorial experiment is being planned for this investigation, the six factors being: Task (go forward vs go backward), Color (blue vs white), Size (same vs different), Text Label Forward (Next vs Save and Continue), Text Label Backward (Back vs Previous), and Symbol Label (includes arrows vs no arrows).

Highlights: During FY2021, staff provided consulting help to the staff at CBSM in the design and eventual modeling and analysis of data from the two experiments listed. For the experiment outlined in (i), statistical methodology was discussed in order to assess the effectiveness, efficiency, and satisfaction of using QR codes to access an online survey. Also discussed was the comparison of the usability and survey response quality between two survey access modes: scanning QR codes and entering a URL. The factorial experiment in (ii) has 64 combinations (64 screens), which is too many for any participant to have in a study. Consulting help was provided on reducing the number of screens per participant in the study.

Work on this project has been completed.

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

D. Experiment on the Effect of Branding Visibility: Design and Analysis

Description: An experiment carried out by the Center for Behavioral Science Methods (CBSM) investigated the effect of branding visibility on participants' confidence in government surveys. The following research questions were addressed: (a) Does the placement of the logo affect participants' ability to identify the organization conducting the survey? (b) Does placing the logo on every page make it more likely for participants to see it? (c) What does the logo of the organization conducting the survey mean to participants? and (d) Which placement of logo do participants prefer?

Highlights: During FY2021, staff provided consulting help to the staff at CBSM. The design used in the experiment was a between-subjects design with a single factor having three levels, indicating three possibilities for the placement of the logo in a mobile survey. The participants completed a 6-question survey on a smartphone. The binary response data was analyzed using logistic regression. The findings showed that a majority of participants did not perceive a logo on a mobile survey per se, but more likely perceived a logo in an invitation letter from the Census Bureau. The data analysis also showed that displaying a logo on a survey was preferred by most participants with the majority preferring the logo on every screen (compared to the logo on the first screen only).

Work on this project has been completed.

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

1.16 NATIONAL CANCER INSTITUTE

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

A. National Cancer Center Tobacco Use Survey/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 FY2021, staff drafted a new work research plan and proposal to renew this project, and a new Inter Agency Agreement (IAA) to renew this project was approved by NCI and U.S. Census Bureau management staff.

Staff improved work on county-level direct survey-based estimates of population coverages of the two tobacco measures of smoke-free workplace policies and smoke-free home rules. County-level estimates for these two tobacco measures were re-calculated for 3,134 counties across the country. Model-based estimates for population coverage of smoke-free workplace policies and smoke-free home rules were produced for 3,134 U.S. counties for 2014-15 TUS-CPS files. Bayesian modeling through a Markov Chain Monte Carlo simulation was used to produce the final model-based county-level estimates for our draft manuscript “Small Area Estimation of Smoke-free Workplace Polices and Home Rules in U.S. Counties.” This draft co-authored manuscript was published in the *Journal of Nicotine and Tobacco Research.*

Additionally, staff developed SAS and R codes to re-calculate the following smoking estimates: (a) State direct estimates of weighted proportions for five smoking outcomes, (b) County Direct estimates of weighted proportions for five smoking outcomes, (c) State and County design effects estimates, (d) Model based Hierarchical Bayesian estimates for five smoking outcomes using regular state fips codes, and (e) Model based Hierarchical Bayesian estimates for five smoking outcomes using state random effects method.

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

Highlights: During FY2021, staff began collaborative research with Social Economic and Housing Statistics Division and Center for Economic Studies researchers on experimental weighting techniques for accounting for massive nonresponse due to data collection interruptions in the American Community Survey. Staff began reviewing the literature on calibration weighting and developed a preliminary study plan for a simulation study of entropy balancing with varying degrees of magnitude of unit missing data and nature of unit missing data. The proposed work involves understanding properties of the experimental weighting method, with the goal of providing statistically principled evaluation and accuracy assessment measures.

Staff: Darcy Steeg Morris (x33989), 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 FY2021, staff worked with Economic Statistical Methods Division (ESMD) staff on documenting the hierarchical Bayesian imputation model methodology for estimating state-level retail sales based

on data from a third party aggregator. These imputations are 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). With this work as a case study, staff is working with colleagues on a paper illustrating Bayesian multiple imputation hierarchical models in Stan for official estimates about the economy using third party data.

Staff: Darcy Steeg Morris (x33989), Yves Thibaudeau

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 FY2021, staff completed a paper titled "Estimation in exponential family regression based on linked data contaminated by mismatch error" and submitted to "Statistics and Its Interface" journal. In this paper we focus on identifying and correcting false matches in record linkage. It is known that due to linkage error statistical analysis with linked data files may need different modeling than usual ones to avoid bias and misleading results. Several recent papers have studied post-linkage linear regression analysis with the response variable in one file and the covariates in a second file from the perspective of the "Broken Sample Problem" and "Permuted Data." In this paper, we present an extension of this line of research to exponential family response given the assumption of a small to moderate number of mismatches. A method based on observation-specific offsets to account for potential mismatches and $\ell 1$-penalization is proposed, and its statistical properties are discussed. We also present sufficient conditions for the recovery of the correct correspondence between covariates and responses if the regression parameter is known. The proposed approach is compared to established baselines, namely the methods by Lahiri and Larsen and Chambers, both theoretically and empirically based on synthetic and real data. The results indicate that substantial improvements over those methods can be achieved even if only limited information about the linkage process is available. We completed a paper titled "Regression with linked data sets subject to linkage error" and submitted in "WIREs Computational Statistics." In this focus article, we give a comprehensive overview of recent development in methodologies for dealing with linkage errors in regression analysis with linked data sets, with an emphasis on recent approaches and their connection to the so-called "Broken Sample" problem. We also provide an empirical study that illustrates the efficacy of some these proposed methods in different scenarios.

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

B. Entity Resolution and Merging Noisy Databases
Description: Work is underway on the problem of

merging noisy databases to remove duplicate entities (individuals, households, etc.), where typically a unique identifier is not known. This problem in the literature is known as entity resolution or record linkage. Work is undertaken on improved methodology, and scalability, and testing such methods on both synthetic and real data.

Highlights: During FY2021, the Marchant et al. (2020) paper was published in the *Journal of Computational and Graphical Statistics*, where a case study of the 2010 decennial census is presented on the state of Wyoming along with administrative data. Code is publicly available via github at dblink. There are still plans to run dblink on additional states and release this in future work.

Staff: Rebecca C. Steorts (919-485-9415), David Brown (CES), Casey Blalock (CODS), Yves Thibaudeau

C. 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 FY2021, Gentles, Reddy, and Steorts (2021) shared an investigation with comparisons of Marchant et al. (2021), Sadinle (2018), and Edmorando et al. (2018) on RLdata500, RLdata10000, and the NLTCS. Code is still under development at this time regarding the comparison of the methods.

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

D. 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 FY2021, staff worked to revise Binette and Steorts (2021) which is a review paper on record linkage/entity resolution. A pre-print of the paper can be found at: https://arxiv.org/abs/2008.04443

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

E. Analysis of Alternatives and Proof of Concept for Record Linkage Modernization NEW
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 FY2021, staff has been supporting requirement professionals by helping to elicit hundreds

of requirements that will be the basis for the enterprise record-linkage solution at the Census Bureau in the near and extended future.

At the 2021 Symposium on Data Science and Statistics, staff presented new approaches to estimate the likelihood of the Fellegi-Sunter model in high-dimensional spaces based on the theory of Fienberg and Rinaldo (2012). Journal submission(s) are in preparation.

Staff: Yves Thibaudeau (x31706), Daniel Weinberg, Jaya Damineni (CODS), Anup Mather (ADRM), Steve Nesbitt (CTR), Monique Latrice Edmonds (CTR), Pushpinder Multani (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 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 FY2021, staff finalized a draft visualization for comparing one state with each of the other states for 88+ different American Community Survey (ACS) topics for each of the years 2018 and 2019. Comparison results using three different variations (significance levels are unadjusted, adjusted with independence, or adjusted with Bonferroni) are presented. Independence or Bonferroni approaches have the advantage of attempting to minimize overreporting of statistically significant differences. The visualization is simple and easy to use, and it is based on Wright, Klein, and Wieczorek (2019) and software (Wieczorek, CRAN). There is a little variation in a tiny fraction of the hundreds of thousands of tests when compared to those published by the ACS which seem to be due to rounding. Development has begun on visualizing uncertainty in an estimated overall ranking of K different populations based on Klein, Wright, and Wieczorek (2020) and software (Wieczorek, CRAN).

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 FY2021, staff worked to produce a draft paper which generalizes results.

Staff: Tommy Wright (x31702)

Apportionment

Highlights: During FY2021, a short article showing the complete details of the 2020 apportionment computations was prepared and appeared in the July 2021 issue of *AMSTATNEWS,* newsletter of the American Statistical Association.

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 FY2021, there was no new activity on this project.

Staff: Eric Slud (x34991)

E. 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 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 FY2021, staff assisted the Center for Behavioral Science Methods (CBSM) - led research into assessment of accuracy of the probability (RDD) and nonprobability (web-panel-based) 2019-2020 Tracking Surveys on attitudes to the decennial census. That assistance took the form of regular meetings with the CBSM staff tasked with assessing the quality of the surveys, and additional methodology development. Specific staff activities included the calculation of weight-adjustment by several different methods, for purposes of comparison with weight-adjustments and estimates produced by the contractor (Team Y&R) which conducted the surveys. Staff developed methods, some of them novel, for handling missing values in poststratification variables ad variables used (in the RDD survey) for the construction of base-weights, and for calculating variances of the demographic and benchmark estimates of population proportions from the surveys. The research results on weighting methodology for these surveys was written into a draft technical report that is being prepared for the Center for Statistical Research and Methodology *Research Report Series* and for later journal publication. The same material was also included in a conference talk to be delivered in October 2022.

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

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 FY2021, staff investigated moment based small area estimation methods that do not rely on distributional assumptions for the random effects. Staff generalized the Fay-Harriot model framework to include non-linear mean functions for count and rate outcomes and address auxiliary variables with measurement error. For the non-normal Fay-Herriot model, staff considered a broad class of estimating equations that include Prasad-Rao, Fay-Herriot, maximum likelihood and Jiang, Nguyen and Rao's overall best prediction estimation methods of the model parameters. Staff developed a second-order accurate approximation to the MSE of the empirical best linear unbiased predictors that result from these model parameter estimation methods. Using this approximation, the staff derived an analytical estimator of the MSE. Staff previously proposed a bootstrap method to accurately estimate this MSE based on

moment-matching method. Staff conducted simulations to investigate the performance of this approach and produced comparisons to other existing methods. Staff showed that the bootstrap approach performed similarly to existing approaches when the sampling variance was small relative to the random effects variance, and that the bootstrap approach was generally less biased when the sampling variance was large. Some preliminary results from this project were presented in September at the Small Area Conference in Naples, Italy.

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 FY2021, there was no reportable 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 FY2021, staff and an external collaborator continued to revise a manuscript prepared in FY2020. This manuscript explores effectiveness of various spatial random effects models as alternative to the Fay-Herriot model. We expanded assessment tools of the effectiveness of these spatial models based on a simulation study and a real application. The revised manuscript was submitted to a journal for publication.

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 FY2021 and as an alternative to the solutions suggested by these authors, staff is pursuing a Bayesian approach. By suitably standardizing the ETMSE, staff constructed a pseudo-likelihood for the model parameters and use a class of noninformative priors to derive Bayesian estimates of small area means. Staff is comparing the proposed pseudo-Bayes method with the frequentist compromised regression weights methods through an example of estimation median incomes of U.S. states. Application and a simulation study show that our Bayesian solution competes favorably with the frequentist methods when assessed based on suitable frequentist criteria. It is claimed in the literature that these alternative methods possess better robustness to model misspecification than the traditional EBLUP methods Staff is investigating to what extent such claim holds. Preliminary results are less supportive of the claim of this robustness of the compromise regression weight estimators of small area means.

Staff: Gauri Datta (x33426)

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 FY2021, 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 chapters on filter constraints, model-based filters, and co-integration; (b) refined modeling, code, and vignettes for Ecce Signum software product, for multivariate seasonal adjustment with missing values; (c) revised multivariate forecasting and nowcasting methodology to generate preliminary releases of transportation index time series; (d) devised new method of computing multivariate autocovariances based on the frequency domain.

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 FY2021, staff made progress on several projects, including (a) revisions on a paper that describes estimation of multivariate time series models using the Frobenius norm; (b) refining the methodology for model identification based on testing for zeroes in nonparametric estimates of the spectral density; (c) derived the quadratic forecast filter, expressing the formula in terms of 2- and 4-dimensional covariance arrays; (d) derived new results on the factorization of polyspectra, involving the use of group representations of symmetries and the cepstral mapping; (e) obtained new asymptotic results for polyspectral means, a type of estimator that involves a weighted integral of the polyspectral density; (f) continued research into methods for multivariate count time series, and completed revisions on a paper describing these results; (g) continued research into multivariate business cycle models, and multivariate trend filters; (h) tested new procedure for local spectral density estimation that is optimal at the boundary of the frequency domain; (i) refined methodology and code for a clustering technique, useful for building implicit networks among covariates; (j) developed code and methodology for testing for slope homogeneity in panel regressions; (k) developed time series differential privacy notions of utility and privacy, and the use of all-pass filtering as a protective device; (l) refined variable selection methodology for vector autoregressions, using Granger causality to build and refined models; (m) developed a new expression for the multivariate missing value filter, for the case of multiple missing values; (n) obtained numerical results depicting the bi-spectral density, demonstrating the bi-spectral factorization.

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

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 FY2021, staff made great strides on a variety of related projects, developing (1) a multivariate CMP distribution based on the compounding method, (2) a bivariate CMP distribution based on the trivariate reduction method, (3) a first-order moving average model based on the sum-of-CMPs distribution which contains the first-order Poisson and negative binomial moving average models as special cases, (4) a longitudinal model based on the CMP distribution, and (5) a vignette for analysts to use as a helpful introduction to the COMPoissonReg R package. Further, staff continue to update the COMPoissonReg package, in response to staff-recognized potential advancements and user queries. Research manuscripts associated with Projects (1) and (3) were published in peer-reviewed journals (Project (1) in the journal *Stats*, and Project (3) in the *Journal of Statistical Distributions and Applications*), while manuscripts associated with Projects (2) and (4) are under review (Staff revised and resubmitted the Project (2) manuscript in response to reviewer feedback).

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

B. Design and Analysis of Embedded Experiments
Description: This project is intended 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 FY2021, staff revised and resubmitted manuscript after receiving feedback from initial journal submission. This manuscript concerns sample size determination under fixed effects continuation-ratio logit models and is connected to Project E "Experiment for Effectiveness of Bilingual Training" in Section 1 of this report.

Staff: Andrew Raim (x37894), Thomas Mathew, 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 FY2021, staff and collaborators completed a draft of research paper on predicting census response rates via interpretable nonparametric additive models with structured interactions. In the paper, we present sparse nonparametric additive models with pairwise interactions as alternative interpretable statistical methods to inefficient linear regression or uninterpretable black-box machine learning methods for predicting response rates in surveys. We implement the proposed methods on the U.S. Census Planning Database and demonstrate that these lead to high-quality predictive

models that permit actionable interpret-ability for different segments of the population. The submitted paper is currently under revision.

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

D. The Difference-in-means Estimator and the Estimation of its Variance in a Finite Population

Description: For comparing two treatments in a finite population setting, the difference-in-means estimator is widely used. However, its variance cannot be estimated since the variance expression involves observations on both the treatments from the same population unit. Available literature suggests conservative estimation of the variance or recommends the use of appropriate bounds on the non-estimable part of the variance; the latter is based on covariate information. The goal of this project is to develop an estimate of the non-estimable part using covariate information, and also to derive new bounds for the non-estimable part. The problem will also be addressed when the difference-in-means estimator is adjusted using covariates. It is anticipated that the proposed research will lead to methodologies that can be applied to the analysis of some of the embedded experiments carried out at the Census Bureau.

Highlights: During FY2021, this project was started during quarter 3. Staff made progress in terms of deriving both an estimate and new bounds for the non-estimable part of the variance of the difference-in-means estimator. Staff is currently looking for an appropriate data set that can be used to illustrate the computation of the estimates making use of covariate information, and the computation of its estimated variance using different machine learning predictive models such as generalized additive, gradient boosting and random forest models. It is hoped that some datasets from ACS will be available for this purpose. In addition, staff is in the planning stages to prepare a manuscript that summarizes the current literature on the topic.

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 FY2021, this project was started during quarter 3. Staff developed a framework for investigating the distribution of the counts based on a sanitized version of the original data, obtained using the Laplace mechanism. An obvious question of interest is that of unbiasedness; that is, whether the means of the resulting counts in each interval are equal to the corresponding counts based on the original data. It appears that the privacy-loss budget has to be appropriately distributed among the different intervals in order to achieve unbiasedness. Staff developed a system of non-linear equations whose solutions will dictate the manner in which the privacy-loss budget has to be distributed. Staff is currently investigating the numerical solution of the system of non-linear equations.

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 FY2021, staff proposed a step function to approximate an internal density needed by the direct sampler. This facilitates proof of an upper bound on the accuracy of the approximation, a rejection sampling method with an upper bound on the probability of a rejection, and a knot-selection algorithm for reducing both bounds efficiently. A manuscript based on this work was submitted for peer review. A reference implementation of the sampler in R and C++ was also provided on Github. Staff began work on an application of the direct sampler to a Sufficient Statistic Perturbation (SSP) setting from Bernstein and Sheldon (NeurIPS, 2018); this is a simplified but holistic differentially private workflow which assures the sensitivity of released statistics is finite.

Staff: Andrew Raim (x37894)

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 FY2021, staff worked on developing randomized response methods for privacy and data confidentiality protection. Staff published two papers, titled "A Post-randomization Method for Rigorous Identification Risk Control in Releasing Microdata" in the *Journal of Statistical Theory and Practice* and "Minimax Randomized Response Methods for Protecting Respondent's Privacy" in *Communications in Statistics - Theory and Methods.* These two papers present novel and optimal randomized response methods for rigorous privacy and data confidentiality protection. Staff also completed a technical report titled "A Review of Rigorous Randomized Response Methods for Protecting Respondent's Privacy and Data Confidentiality." This article has been accepted for publication in an edited volume in honor of Prof. C.R. Rao.

Staff: Tapan Nayak (x35191)

B. Bayesian Analysis of Singly Imputed Synthetic Data
Description: Under this project, staff members will conduct research on some aspects of Bayesian analysis of singly 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 FY2021, staff had a paper accepted for publication: Guin, A., Roy, A., and Sinha, B. (In Press). "Bayesian Analysis of Singly Imputed Partially Synthetic Data Generated by Plug-In Sampling and Posterior Predictive Sampling under the Multiple Linear Regression Model," *International Journal of Statistical*

Applications.

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

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

Highlights: During FY2021, staff worked on the paper "Bayesian Analysis of Multiply Imputed Synthetic Data under Multiple Linear Regression Model" (Abhishek Guin, Anindya Roy, and Bimal Sinha) which is still being finalized and will be submitted to a journal early in FY2022.

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

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

Highlights: During FY2021, staff worked on a technical report based on Bayesian analysis of singly imputed synthetic data under a multivariate normal model is still being prepared.

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

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 FY2021, staff organized the twelfth annual *2021 SUMMER AT CENSUS* which brought 13 recognized scholars to the Census Bureau for 1-3 day (virtual) visits covering broad topic themes including: demography, disclosure avoidance, funds allocation, health sample surveys, sampling and estimation, small area estimation, social media analysis, and survey methodology. Each scholar engaged in collaborative research with Census Bureau researchers and staff (Center for Statistical Research and Methodology, Center for Behavioral Science Methods, Decennial Statistical Studies Division, Economic Statistical Methods Division, Communications Directorate, Center for Economic Studies, Center for Enterprise Dissemination, Population Division, and Research and Methodology Directorate) 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

Betancourt, B., Zanella, G., and Steorts, R. (In Press). "Random Partition Models for Microclustering Tasks," *Journal of the American Statistical Association, Theory and Methods.*

Binder, C., McElroy, T., and Sheng, X. (In Press). "Term Structure of Uncertainty: New Evidence from Survey Expectations," Published online, *Journal of Money, Credit, and Banking*.

Chai, J. and Nayak, T.K. (2021). "Minimax Randomized Response Methods for Protecting Respondent's Privacy," *Communications in Statistics - Theory and Methods*, https://doi.org/10.1080/03610926.2021.1973503

Chen, B., McElroy, T., and Pang, O. (In Press). "Assessing Residual Seasonality in the U.S. National Income and Product Accounts Aggregates," *Journal of Official Statistics.*

Davis, R.A., Fokianos, K., Holan, S., Joe, H., Livsey, J., Lund, R.B., Pipiras, V., and Ravishanker, N. (In Press). "Count Time Series: A Methodological Review," *Journal of the American Statistical Association*.

Feng, X., Mathew, T, and Adragni, K. (2021). "Interval Estimation of the Intra-class Correlation in General Linear Mixed Effects Models," *Journal of Statistical Theory and Practice*, *15, Article 65*.

Franco, C. and Bell, W.R. (In Press). "Using American Community Survey Data to Improve Estimates from Smaller U.S. Surveys through Bivariate Small Area Estimation Models," *Journal of Survey Statistics and Methodology*.

Guin, A., Roy, A., and Sinha, B. (In Press). "Bayesian Analysis of Singly Imputed Partially Synthetic Data Generated by Plug-In Sampling and Posterior Predictive Sampling under the Multiple Linear Regression Model," *International Journal of Statistical Applications.*

Janicki, R., Raim, A.M., Holan, S.H., and Maples, J. (In Press). "Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models with Application to American Community Survey Special Tabulations," *The Annals of Applied Statistics.*

Jia, Y., Kechagias, S., Livsey, J., Lund, R., Pipiras, V. (2021). "Latent Gaussian Count Time Series Modelling," *Journal of the American Statistical Association*.

Liu, B., Dompreh, I., and Hartman, A.M. (2021). "Small Area Estimation of Smoke-Free Workplace Policies and Home Rules in U.S. Counties," *Journal of Nicotine and Tobacco Research*.

Marchant, N., Kaplan, A., Rubenstein, B., Elzar, D., and Steorts, R. (2021). "d-blink: Distributed End-to-End Bayesian Entity Resolution," *Journal of Computational Graphics and Statistics, 30(2),* 406-421.

McElroy, T. (2021). "A Diagnostic for Seasonality Based upon Polynomial Roots of ARMA Models," *Journal of Official Statistics, 37(2)*, 1-28.

McElroy, T. (In Press). "Frequency Domain Calculation of Seasonal VARMA Autocovariances," *Journal of Computational and Graphical Statistics*.

McElroy, T. and Das, S. (2021). "Nonlinear Prediction via Hermite Transformation," *Statistical Theory and Related Fields 5(1)*, 49-54.

McElroy, T. and Roy, A. (2021). "Testing for Adequacy of Seasonal Adjustment in the Frequency Domain," *Journal of Statistical Planning and Inference*, *221*, 241-255.

McElroy, T. and Roy, A. (In Press). "Model Identification via Total Frobenius Norm of Multivariate Spectra," *Journal of the Royal Statistical Society, Series B*.

McElroy, T., Roy, A., Livsey, J., Firestine, T., and Notis, K. (2021). "Anticipating Revisions in the Transportation Services Index," *Journal of the International Association of Official Statistics*, *37*, 641-653.

Mosaferi, S., Ghosh, M., and Steorts, R. (In Press). "Measurement Error Models for Small Area Estimation," *Communications and Statistics: Simulation and Computation.*

Moura, R., Klein, M., Zylstra, J., Coelho, C., and Sinha, B. (In Press). "Inference for Multivariate Regression Model Based on Synthetic Data Generated Under Plug-In Sampling," *Journal of the American Statistical Association (Theory & Methods)*.

Mulry, M., Bates, N., and Virgile, M. (2021). "Viewing Participation in Censuses and Surveys through the Lens of Lifestyle Segments" (print), *Journal of Survey Statistics and Methodology*, doi:1093/jssam/smaa006.

Parker, P.A., Holan, S.H., and Janicki, R. (In Press). "Computationally Efficient Bayesian Unit-Level Models for Non-Gaussian Data Under Informative Sampling," *The Annals of Applied Statistics*.

Sellers, K.F., Arab, A., Melville, S., and Cui, F. (2021). "A Flexible Univariate Moving Average Time-Series Model for Dispersed Count Data," *Journal of Statistical Distributions and Applications 8 (1).* https://doi.org/10.1186/s40488-021-00115-2

Sellers, K.F., Li, T., Wu, Y., and Balakrishnan, N. (2021). "A Flexible Multivariate Distribution for Correlated Count Data," *Stats*, *4(2),* 308-326, https://doi.org/10.3390/stats4020021.

Slawski, M., Diao, G., and Ben-David, E. (2021). "A Pseudo-Likelihood Approach to Linear Regression with Partially Shuffled Data," *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2020.1870482

Trimbur, T. and McElroy, T. (In Press). "Modelled Approximations to the Ideal Filter with Application to GDP and its Components," *The Annals of Applied Statistics.*

Wang, Z., Ben-David, E., Diao, G., and Slawski, M. (2021). "Regression with Linked Datasets Subject to Linkage Error," Wiley Interdisciplinary Reviews: Computational Statistics, DOI: 10.1002/wics.1570

Wright, T. (2020). "A General Exact Optimal Sample Allocation Algorithm: With Bounded Cost and Bounded Sample Sizes," *Statistics and Probability Letters, Vol 165*, Article 108829.

Zhai, X., and Nayak, T.K. (2021). "A Post-randomization Method for Rigorous Identification Risk Control in Releasing Microdata," *Journal of Statistical Theory and Practice*, 15, Article 8, https://doi.org/10.1007/s42519-020-00143-2.

Zhao, J., Mathew, T., and Bebu, I. (2021). "Accurate Confidence Intervals for Inter-Laboratory Calibration and Common Mean Estimation," *Chemometrics and Intelligent Laboratory Systems*, *208*. DOI: 10.1016/j.chemolab.2020.104218.

Zimmer, Z., Park, D., and Mathew, T. (2021). "Tolerance Limits under Zero-Inflated Lognormal and Gamma Distributions," *Computational and Mathematical Methods*, *Special Issue on Statistics, 3*. DOI: 10.1002/cmm4.1113.

3.2 BOOKS/BOOK CHAPTERS

Erciulescu, A., Franco, C., and Lahiri, P. (2021). "Use of Administrative Records in Small Area Estimation," in Chun, A. Y. and Larsen, M. (Eds.), *Administrative Records for Survey Methodology,* New York, NY: Wiley Publishers.

Parker, P.A., Janicki, R., and Holan, S. (In Press). "Bayesian Methods Applied to Small Area Estimation for Establishment Statistics," in Bavdaž, M., Bender, S., Jones, J., MacFeely, S., Sakshaug, J.W., Thompson, K.J., and van Delden, A. (Eds.), *Advances in Business Statistics, Methods and Data Collection*, Wiley.

Thibaudeau, Y., Slud, E., and Cheng, Y. (2021). "Small-Area Estimation of Cross-Classified Gross Flows Using Longitudinal Survey Data," *Advances in Longitudinal Survey Methodology,*469-489, Peter Lynn ed., Wiley.

3.3 PROCEEDINGS PAPERS

Joint Statistical Meetings, American Statistical Association, (Virtual, USA), August 2-6, 2020
2020 Proceedings of the American Statistical Association

- Mary Mulry, Steven Scheid, Darcy Morris, and Nancy Bates, "Evaluation of Multi-class Classification Models for Census Mindsets," 2306-2319.
- Eric Slud, "Nonidentifiability of Mixed-Model Parameters under Informative Sampling Using Only Single-Inclusion Weights," 93-106.

3.4 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY RESEARCH REPORTS

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

RR (Statistics #2020-06): Tapan K. Nayak, "A Review of Rigorous Randomized Response Methods for Protecting Respondent's Privacy and Data Confidentiality," October 6, 2020.

RR (Statistics #2020-07): Hee Cheol Chung and Gauri Sankar Datta, "Bayesian Hierarchical Spatial Models for Small Area Estimation," November 19, 2020.

RR (Statistics #2021-01): Andrew M. Raim, "Direct Sampling in Bayesian Regression Models with Additive Disclosure Avoidance Noise," March 15, 2021.

RR (Statistics #2021-02): Abhishek Guin, Anindya Roy, and Bimal Sinha, "Bayesian Analysis of Singly Imputed Partially Synthetic Data Generated by Plug-in Sampling and Posterior Predictive Sampling under the Multiple Linear Regression Model," August 25, 2021.

3.5 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY STUDY SERIES

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

SS (Statistics #2021-01): Tommy Wright and Kyle Irimata, "Empirical Study of Two Aspects of the TopDown Algorithm Output for Redistricting: Reliability & Variability," May 28, 2021.

SS (Statistics #2021-02): Tommy Wright and Kyle Irimata, "Empirical Study of Two Aspects of the TopDown Algorithm Output for Redistricting: Reliability & Variability (August 5, 2021 Update)," August 5, 2021.

3.6 OTHER REPORTS

Mulry, M. H., Mule, T., Keller, A. K., and Konicki, S. (2021). "Overview of Administrative Records Modeling in the 2020 Census." *2020 CENSUS PROGRAM MEMORANDUM SERIES: 2021.10*. https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/administrative-record-modeling-in-the-2020-census.pdf

Wright, T. (2021). "Demystifying Apportionment Computations for the U.S. House of Representatives," *AMSTATNEWS, July*, American Statistical Association, Alexandria, VA, 8-11.

Benjamini, Y., D. De Veaux, R., Efron, B., Evans, S., Glickman, M., Graubard, B.I., He, X., Meng, X.-L., Reid, N., Stigler, S.M., Vardeman, S.B., Wikle, C.K., Wright, T., Young, L.J., and Kafadar, K. (2021). "The ASA President's Task Force Statement on Statistical Significance and Replicability," *The Annals of Applied Statistics, 15(3),* 1084-1085. DOI: 10.1214/21-AOAS1501

[*Excerpt*: "…the use of P-values and significance testing, properly applied and interpreted, are important tools that should not be abandoned… P-values are valid statistical measures that provide convenient conventions for communicating the uncertainty inherent in quantitative results. Indeed, P-values and significance tests are among the most studied and best understood statistical procedures in the statistics literature. They are important tools that have advanced science through their proper application…"]

4. TALKS AND PRESENTATIONS

Mathematics and Statistics Departmental Seminar (virtual), Slippery Rock University, Slippery Rock, PA, November 5, 2020.

- Kimberly F. Sellers, "Flexible Regression Models for Dispersed Count Data."

Government Advances in Statistical Programming (GASP) 2020, U.S. Bureau of the Census, Washington, D.C., November 6, 2020.

- Yves Thibaudeau, "Toward an Integrated Environment in Record Linkage."

UMBC Statistics Colloquium, University of Maryland at Baltimore County, Virtual, November 13, 2020.

- Darcy Morris, "A Conway-Maxwell-Multinomial Distribution for Flexible Modeling of Clustered Categorical Data."

DC R Conference: The Government & Public Sector (virtual), Washington, D.C., December 4, 2020.

- Kimberly F. Sellers, "Analyzing Count Data Expressing Data Dispersion."

UMBC Statistics Colloquium, Baltimore, Maryland, December 11, 2020.

- Ryan Janicki, "Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models with Application to American Community Survey Special Tabulations."

International Virtual Conference on Advanced Statistical Techniques in Business and Industry, A Regional Virtual Conference of International Society for Business and Industrial Statistics (ISBIS) In Conjunction with Silver Jubilee Anniversary of Department of Statistics, Cochin University of Science and Technology (CUSAT), December 28-30, 2020.

- James Livsey, "Multivariate Signal Extraction with Latent Component Models."

Department of Bioinformatics and Biostatistics Seminar Series, University of Louisville, Louisville, KY, January 22, 2021.

- Rebecca Steorts, "(Almost) All of Entity Resolution."

Seminar (virtual), Data Science Institute, University of California, San Diego, CA, March 17, 2021.

- Tucker McElroy, "Polyspectral Factorization."

United States Patent and Trademark Office (USPTO) Symposium on Entity Resolution (Invited Talk & Invited Panel Discussion), Alexandria, VA, March 22, 2021.

- Rebecca Steorts, "(Almost) All of Entity Resolution."

Research Interaction Team Seminar Series (virtual), Joint Program on Survey Methodology, University of Maryland at College Park, College Park, MD, March 30, 2021

- Carolina Franco, "Small Area Estimation in the Presence of Measurement Error," March 30, 2021.

University of Kentucky Statistics Seminar Series (virtual), Lexington, KY, April 2, 2021.

- Andrew Raim, "Direct Sampling in Bayesian Regression Models with Additive Disclosure Avoidance Noise."

American Statistical Association Webinar, April 8, 15, 22, 29, 2021.

- Tucker McElroy, "Time Series: A First Course with Bootstrap Starter."

Research Interaction Team on Small Area Estimation, April 20, 2021.

- Eric Slud, "Hybrid BRR and Parametric-Model Variance Estimates for Small Domains in Large Surveys."

Joint Seminar (virtual) – Department of Statistics, and Department of Biostatistics, Northwestern University, April 28, 2021.

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

Federal Forecasters Conference, Washington, D.C., May 6, 2021.

- Tucker McElroy, "Ecce Signum: An R Software Package for Analyzing Multivariate Time Series."

Department of Mathematics and Computer Science, Faculty Seminar Series, Davidson College, Davidson, NC., May 10, 2021.

- Rebecca Steorts, "(Almost) All of Entity Resolution."

Research Interaction Team Seminar Series (virtual), Joint Program on Survey Methodology, University of Maryland at College Park, College Park, MD, May 11, 2021.

- Carolina Franco, Panelist, "United Nation's Toolkit on Small Area Estimation for Sustainable Development Goals."

Statistics Seminar (virtual), Department of Statistics, University of California, Davis, Davis, CA, May 13, 2021.

- Carolina Franco, "Small Area Estimation in the Presence of Measurement Error in the Covariates."

International Indian Statistical Association Conference (virtual) 2021, May 21, 2021.

- Gauri Datta, "An Ad-hoc Pseudo-Bayes Small Area Estimation via Compromise Regression Weights."

Statistical Society of Canada Annual Meeting (virtual) 2021, June 11, 2021.

- Gauri Datta, "An Ad-hoc Pseudo-Bayes Small Area Estimation via Compromise Regression Weights."

Sixth International Conference on Establishment Statistics (ICES VI), Virtual, June 14-21, 2021.

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

Summer Program in Research and Learning (SPIRAL), American University, June 17, 2021 (virtual).

- Kimberly Sellers, "Flexible Regression Models for Dispersed Count Data."

Washington Statistical Society President's Invited Lecture and NASS Research and Development Division Seminar Series, Virtual Seminars, June 24, 2021.

- Rebecca Steorts, "Entity Resolution with Societal Impacts in Statistical Survey Methodology, Statistical Science, and Machine Learning."

Webinar on Poverty Mapping Using Small Area Estimation Techniques (virtual), Organized jointly by the United Nation's Economic Commission for Latin America and the Caribbean (ECLAC) and the United Nation's Inter-Secretariat Working Group on Household Surveys, July 1, 2021.

- Carolina Franco, "SAIPE: Small Area Estimation in the United States."

ISBA International World Meeting, Virtual Short Course, July 3, 2021.

- Rebecca Steorts, "Some of Bayesian Entity Resolution."

2021 World Congress of Statistics (virtual), July 11-16, 2021.

- Tucker McElroy, "Treatment of Time Series Outliers via Maximum Entropy."

2021 MAA-NAM David Harold Blackwell Lecture, MAA MathFest 2021 (Virtual), Mathematical Association of America/National Association of Mathematicians, August 5, 2021.

- Tommy Wright, "2020 Census, Lagrange's Identity, and Apportionment of the U.S. House of Representatives."

2021 Joint Statistical (virtual) Meetings, American Statistical Association, August 8-12, 2021.

- Gauri Datta, "Bayesian Hierarchical Spatial Models for Small Area Estimation."
- 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."
- Mary Mulry, "Overview of Administrative Records Modeling in the 2020 Census."
- Kimberly F. Sellers, "A Flexible Univariate Moving Average Time-Series Model for Dispersed Count Data."

BIG4small: Small Area Estimation 2021 Conference on Big Data and Small Area Estimation (virtual), September 20-24, 2021.

- Gauri Datta, "Error-in-Covariates in Small Area Estimation and a Generalized Fay-Herriot Model."

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

Andrew Raim, U.S. Bureau of the Census, "Direct Sampling in Bayesian Regression Models with Additive Disclosure Avoidance Noise," April 6, 2021.

Tucker McElroy, U.S. Bureau of the Census, "Polyspectral Factorization," April 20, 2021.

Jerzy Wieczorek, Colby College, *SUMMER (Virtually) AT CENSUS*, "K-Fold Cross-Validation for Complex Sample Surveys," May 24, 2021.

Zeina Mneimneh, University of Michigan, *SUMMER (Virtually) AT CENSUS*, "Comprehensive Evaluation of Consent to Link Twitter Data to Survey Data," May 26, 2021.

Partha Lahiri, University of Maryland, College Park, *SUMMER (Virtually) AT CENSUS*, "Estimation of COVID-19 Vaccine Hesitancy Indicators for Small Areas using Statistical Linkage of Multiple Disparate Data Sources," June 2, 2021.

Aleksandra Slavkovic, Pennsylvania State University, *SUMMER (Virtually) AT CENSUS*, "Differentially Private Heavy-Tailed Synthetic Data," June 2, 2021.

Andrew Reamer, The George Washington University, *SUMMER (Virtually) AT CENSUS*, "Understanding the Role of the Decennial Census in the Geographic Allocation of Federal Funding: An Overview of Methods," June 17, 2021.

Jan Vink, Cornell University, *SUMMER (Virtually) AT CENSUS*, "Dimensions of Estimation Errors and Error Evaluation Metrics," June 23, 2021.

Karen Bogen, Mathematica Policy Research, *SUMMER (Virtually) AT CENSUS,* "Surveys as an Essential Tool in the Health Program Evaluation Toolbox," July 1, 2021.

Kenneth Johnson, The University of New Hampshire, *SUMMER (Virtually) AT CENSUS,* "Population Redistribution Trends along the Rural-Urban Continuum, 2000 to 2019," August 2, 2021.

Laura Wilson, Office for National Statistics, United Kingdom, *SUMMER (Virtually) AT CENSUS,* "Respondent Centered Surveys: Theory," August 4, 2021.

Emma Dickinson, Office for National Statistics, United Kingdom, *SUMMER (Virtually) AT CENSUS,* "Respondent Centered Surveys: In-Practice Case Studies," August 4, 2021.

Johann Gagnon-Bartsch, University of Michigan, *SUMMER (Virtually) AT CENSUS,* "Qualitative Insights from Social Media Data," August 11, 2021.

Frederick Conrad, University of Michigan and Michael Schober, The New School, *SUMMER (Virtually) AT CENSUS,* "Better Understanding When and How Social Media Posts Can Augment Public Opinion Surveys," August 12, 2021.

Tommy Wright, U.S. Bureau of the Census, "Apportionment," September 14, 2021.

Benjamin Leinwand, University of North Carolina, Chapel Hill, "Block Dense Weighted Networks with Augmented Degree Correction," September 30, 2021.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

American Statistical Association's 2021 W.J. Youden Award in Interlaboratory Testing

- **Thomas Mathew -** The award is based on the paper (jointly with J. Zhao and I. Bebu) "Accurate Confidence Intervals for Inter-laboratory Calibration and Common Mean Estimation," which appeared in the journal *Chemometrics and Intelligent Laboratory Systems* (2021)

American Statistical Association Fellow

- **Kimberly Sellers -** For excellence in dispersed count data methodology; impactful collaborations with the U.S. Census Bureau; and exemplary leadership in increasing justice, equity, diversity, and inclusion in the statistics and broader mathematical sciences communities.

6.2 SIGNIFICANT SERVICE TO PROFESSION

Emanuel Ben-David

- Refereed a paper for *Survey Methodology*
- Member, Ph.D. Defense Committee, Department of Statistics, George Mason University
- Member, Ph.D. Defense Committee, Department of Mathematics and Statistics, University of Maryland, Baltimore County
- Reviewed papers for *Mathematical Reviews* and *Survey Methodology*
- Committee Member and Reviewer, 2021 SBP-BRiMS (2021 International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation)

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, Two Invited Sessions, International Indian Statistical Association Conference 2021: 1. New Developments in Statistics, 2. Small Area Estimation: New Methods
- Chair, Session (Supersaturated and Sequential Designs), International Indian Statistical Association Conference

Carolina Franco

- Associate Editor, *Journal of the Royal Statistical Society-Series A*
- Chair, Committee on International Relations in Statistics, American Statistical Association

Kyle Irimata

- Refereed papers for *Survey Methodology*, *Statistical Methods in Medical Research* and *Journal of the Royal Statistical Society-Series B*

Ryan Janicki

- Refereed papers for *Bayesian Analysis, International Statistical Review, Journal of the Royal Statistical Society (Series A), Journal of Statistical Software,* and *Journal of Survey Statistics and Methodology*

Patrick Joyce

- Refereed papers for *Journal of the Royal Statistical Society (Series A)* and *Communications in Statistics-Simulations and Computations*

James Livsey

- Organizer, Invited Session (Modelling of Time Series Data - Challenges and Applications to Economic Statistics), *Sixth International Conference on Establishment Statistics (ICES VI)*

Jerry Maples

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

Thomas Mathew

- Associate Editor, *Sankhya*
- Associate Editor, *Journal of Multivariate Analysis*
- Associate Editor, *Journal of Occupational and Environmental Hygiene*
- Refereed papers for *Journal of the American Statistical Association*, *The American Statistician, and BMC Medical Research Methodology*

Tucker McElroy

- Refereed papers for *Econometric Theory, MDPI Econometrics, IEEE, Journal of the Royal Statistical Society Series B, Sankhya, Journal of Business and Economics Statistics, Journal of Time Series Analysis, Statistical Analysis and Data Mining, and Statistical Papers*
- Program Chair, Business and Economics Statistics Section, American Statistical Association
- Zellner Thesis Award Committee, Business and Economics Statistics Section, American Statistical Association

Darcy Morris

- Associate Editor, *Communications in Statistics*
- Treasurer, Survey Research Methods Section, American Statistical Association
- Newsletter Editor, Survey Research Methods Section, American Statistical Association
- Reviewed papers for *Journal of Multivariate Analysis, Statistical Papers, Public Opinion Quarterly, TEST,* and *The American Statistician*

Mary Mulry

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

Tapan Nayak

- Associate Editor, *Journal of Statistical Theory and Practice*
- Refereed papers for *Statistics & Probability Letters, Communications in Statistics - Theory & Methods and Open Journal of Statistics.*

Ned Porter

- Reviewer, IEEE International Conference on Data Science and Advanced Analytics

Andrew Raim

- Refereed a paper for *The American Statistician*

Kimberly Sellers

- Associate Editor, *The American Statistician*
- Associate Editor, *Journal of Computational and Graphical Statistics*
- Commissioning Editor, *WIREs Computational Statistics*
- Refereed papers for *Austrian Journal of Statistics, Journal of Time Series, Communications in Statistics – Theory and Methods, Computational Statistics and Data Analysis,* and *Biometrical Journal*
- Member, Advisory Board, Summer Program in Research and Learning (SPIRAL), American University
- Member, American Statistical Association (ASA) External Nominations and Awards Committee
- Inaugural chairperson, ASA Justice, Equity, Diversity, and Inclusion (JEDI) Outreach Group

Bimal Sinha

- Associate Editor, *Environmental Modeling and Assessment*

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 *Communications in Statistics- Theory & Methods, Environment, Annals of Statistics, Journal of the American Statistical Association,* and *BMC Medicine*

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*

Tommy Wright

- Associate Editor, *The American Statistician*
- Member, Task Force on Statistical Significance and Replicability, American Statistical Association

6.3 PERSONNEL NOTES

- Carolina Franco accepted a position with the National Opinion Research Center (NORC).

APPENDIX A **Center for Statistical Research and Methodology FY 2021**
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		
6550G01	Data Coding/Editing/Imputation		
6550G06	Redistricting Data Program		
6550G08	Data Products Dissemination Preparation/Review/Approval		
6650G01	PES Planning and Project Management		
6650G20	2020 Evaluations-Planning and Project Management		
6650G25	2030 Planning and Project Management		
6750G01	Administrative Records Data		
	1. Summary of Administrative Records Modeling for Some Enumerations in the 2020 Census	Mary Mulry	Tom Mule
	2. Supplementing and Supporting Non-Response with Administrative Records	Michael Ikeda	Tom Mule
	3. 2020 Census Privacy Variance	James Livsey	Phil Leclerc
	4. Identifying "Good" Administrative Records for 2020 Census NRFU Curtailment Targeting	Darcy Morris	Tom Mule
	5. Experiment for Effectiveness of Bilingual Training	Andrew Raim	Renee Ellis
	6. Record-Linkage Support for the Decennial Census	Dan Weinberg	David Brown
	7. Coverage Measurement Research	Jerry Maples	Tim Kennel
	8. Assessing Variability of Data Treated by TopDown Algorithm for Redistricting	Tommy Wright	John Abowd
	9. Empirical Investigation of the Minimum Total Population of a Geographic District to Have Reliable Characteristics of Various Demographic Groups	Tommy Wright	John Abowd
	10. Statistical Modeling to Augment 2020 Disclosure Avoidance System	Andrew Raim	Michael Walsh
6385G70	American Community Survey (ACS)		
	11. Voting Rights Section 203 Model Evaluation and Enhancements Towards 2021 Determinations	Eric Slud	James Whitehorne
	12. Model-based Estimates to Improve Data Confidentiality for ACS Special Tabulations	Jerry Maples	John Abowd
	DEMOGRAPHIC		
TBA	Demographic Statistical Methods Division (DSMD) Special Projects		
	13. 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		
	14. Data Integration	Edward Porter	Christopher Boniface
7165021	Social, Economic, and Housing Statistics Division Small Area Estimation Projects		
	15. Research for Small Area Income and Poverty Estimates (SAIPE)	Jerry Maples	Wes Basel
	16. Assessing Constant Parameters across Areas in the SAIPE Models	Jerry Maples	Wes Basel
	17. Small Area Health Insurance Estimates (SAHIE)	Ryan Janicki	Wes Basel
	ECONOMIC		
1183X01/ 1183X90/ 2120G90	Economic Statistical Collection		
	18. Use of Big Data for Retail Sales Estimates	Darcy Morris	Stephen Kaputa
	19. Seasonal Adjustment Support	Tucker McElroy	Kathleen McDonald-Johnson
	20. Seasonal Adjustment Software Development and Evaluation	James Livsey	Kathleen McDonald-Johnson
	21. Research on Seasonal Time Series - Modeling & Adjustment Issues	Tucker McElroy	Kathleen McDonald-Johnson
	22. Supporting Documentation & Software for Seasonal Adjustment	James Livsey	Kathleen McDonald-Johnson
	23. Redesign of Economic Sample Surveys (Stratification)	Eric Slud	Justin Z. Smith
	24. Exploring New Seasonal Adjustment & Signal Extraction Methods	James Livsey	Colt Viehdorfer
	25. Classification of Businesses for the NAICS	Emanuel Ben-David	Anne Russell
	26. Production & Dissemination of Economic Indicators	Adam Hall	William Davie
	PROGRAM DIVISION OVERHEAD		
0331000	*27. Research Computing*	Chad Russell	Jaya Damineni
	28. Experiments to Improve Web Surveys: Design & Analysis	Thomas Mathew	Lin Wang

APPENDIX B

FY 2021 PROJECT PERFORMANCE MEASUREMENT QUESTIONNAIRE

CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY

Dear

As a sponsor for the FY 2021 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 2021 ***(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 2021 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 2021 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 (Duke U.)
- Dan Weinberg

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

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
- Ryan Janicki

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

Tucker McElroy (Acting)
- Soumendra Lahiri (Washington U.)
- Anindya Roy (UMBC)
- Thomas Trimbur

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