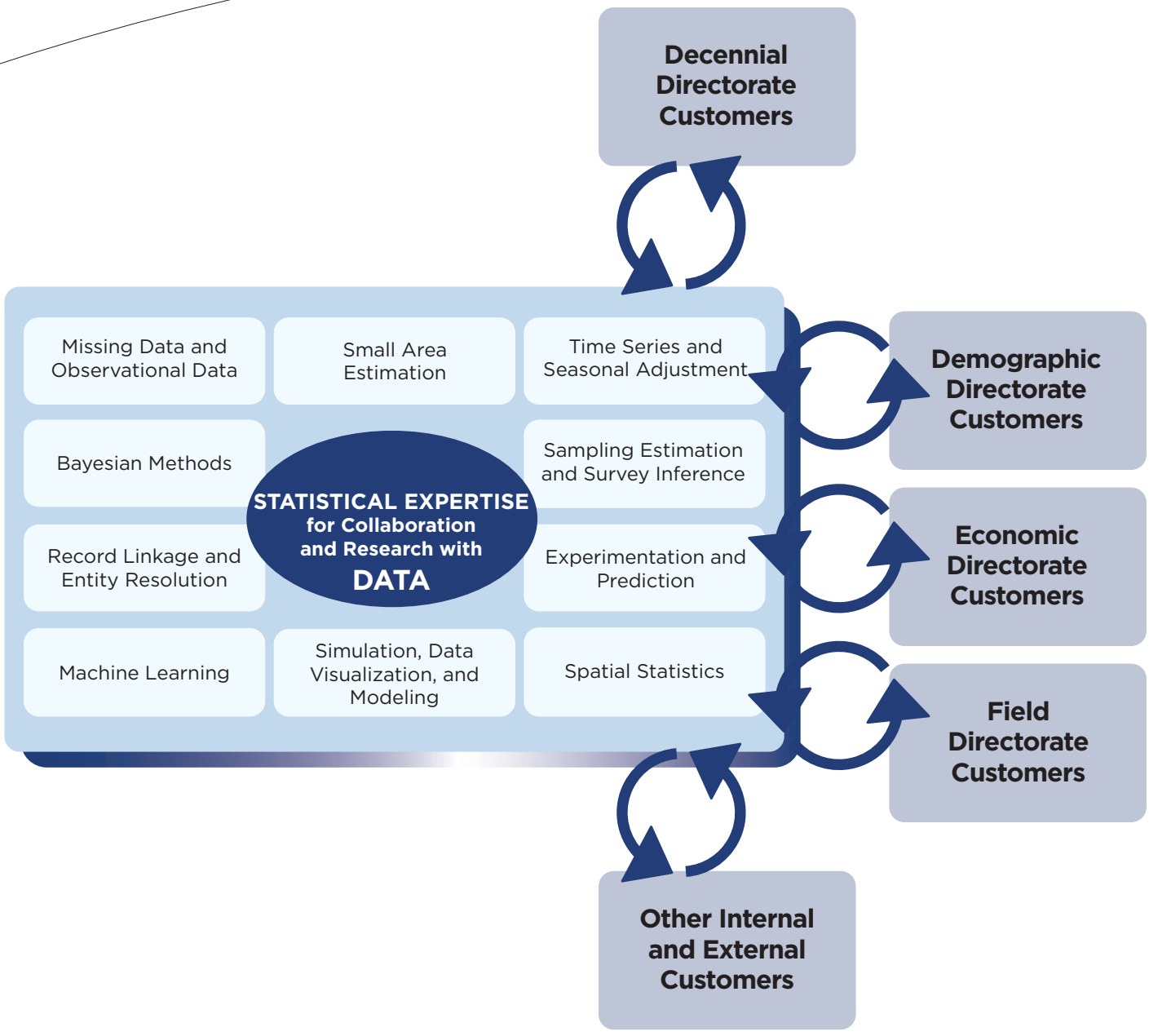
Annual Report of the Center for Statistical Research and Methodology

Research and Methodology Directorate

Fiscal Year 2024



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
4600 Silver Hill Road
Washington, DC 20233
301-763-1702

We help the Census Bureau improve its processes and products. For fiscal year 2024, 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 2024 follow, and more details are provided within subsequent pages of this report:

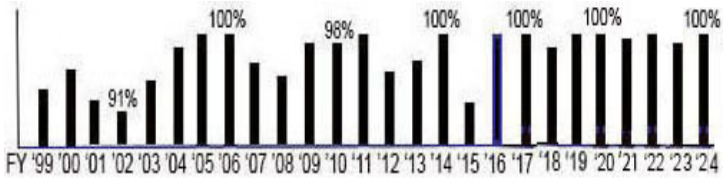
- CSRM researchers and colleagues in the Decennial Statistical Studies Division and Research & Methodology Directorate published results of a study reporting that when an address had both a roster from Administrative Records and a roster from either a self-response or a roster obtained during a Nonresponse Follow-up interview, the response submitted by the household was the one that was used for the census enumeration in most circumstances. [CSRM (Mulry); DSSD (Tello-Trillo, Keller); R&M (Mule)]
- CSRM researchers published a paper documenting the statistical research and development undertaken with the Decennial Statistical Studies Division for the *Section 203, Voting Rights Act,* 2021 Determinations identifying which of nearly 8,000 jurisdictions must provide voting materials in languages in addition to English. [CSRM (Slud, Hall, Kang, Franco); DSSD (Asiala)]
- CSRM researchers and colleagues in the Economic Statistical Methods Division and Economic Indicators Division published a paper on a hierarchical Bayesian mass imputation methodology for estimating state-level retail sales based on data from a third-party aggregate, illustrating the usefulness of Bayesian multiple imputation hierarchical models (and ease of fitting with off-the-shelf software) for official statistics about the economy. [CSRM (Morris); ESMD (Kaputa, Thompson); EID (Hutchinson)]
- CSRM researchers and external collaborators gained publication acceptance of methodology and algorithms, for improved predictions of response rates, that extend the computational frontiers of existing algorithm for sparse additive statistical models showing that in addition to being useful from an interpretability standpoint, the proposed statistical models also lead to predictions that appear to be better than popular black-box machine learning methods. [CSRM (Ben-David); MIT (Ibrahim, Mazumder); University of Sydney, Australia (Radchenko)]
- A CSRM researcher documented results using a new method to match records in the Census Edited File (CEF) to records in the Demographic Frame. The new method uses only PIK instead of both PIK and MAFID. Matching using only PIK sharply increases the proportion of Demographic Frame records that match the CEF. [CSRM (Ikeda)]
- CSRM researchers and a colleague in the Economic Statistical Methods Division finalized a simulation study showing that multivariate signal extraction often outperforms univariate signal extraction using data from the Census Bureau's M3 (manufacturer's shipments, inventories, and orders) Survey. [CSRM (Livsey, Pang); ESMD (Viehdorfer)]
- CSRM researchers published a review article on Bayesian entity resolution, with: (1) a focus on the specific challenges that it poses, (2) a review of prior methods for convergence diagnostics, (3) recommendations for using MCMC sampling, and (4) simulated data, observing that a commonly used Gibbs sampler performs poorly compared to two alternatives. [CSRM (Aleshin-Guendel, Steorts)]
- CSRM researchers and external collaborators published new methodology designed to adjust modern predictive modeling techniques for the presence of mismatch errors in linked data sets; the proposed approach, based on mixture modeling, is general enough to accommodate various prediction modeling techniques in a unified fashion. [CSRM (Ben-David); University of Michigan (West); George Mason University (Slawski)]
- CSRM and Research & Methodology Directorate researchers published results using a Bayesian hierarchical model which utilizes auxiliary publicly available data sources, such as American Community Survey data and past decennial census data. It leverages spatial correlations to improve the precision of the estimates that result from noisy measurements, showing that accurate, model-based estimates of the number of persons in a detailed race-ethnicity group in a target geography, which may be misaligned from the source data, such as an American Indian or Alaska Native (AIAN) area, can be achieved through a novel change-of-support approach. [CSRM (Janicki, Irimata, Livsey, Raim); R&M (Holan)]
- A CSRM researcher and external collaborators published new small area estimation methodology called the observed best prediction (OBP) method which estimates the model parameters by minimizing an objective function that is implied by the total mean squared prediction error. Data analysis and simulation show that the pseudo-Bayesian estimators (PBEs) compete favorably with current methodology. [CSRM (Datta); University of Florida (Lee); University of Georgia (Li)]
- A CSRM researcher and external collaborator published research results on testing the equality of means of several univariate normal populations with a common unknown variance, when the data used for analysis arise from a synthetic version of the original observation. A measure of privacy protection is also evaluated. [CSRM (Sinha); Sister Nivedita University (Basak)]

How Did We[1] Do...

For the 26th year, we received feedback from our sponsors. Near the end of fiscal year 2024, our efforts on 32 of our programs (Decennial, Demographic, Economic, External, etc.) sponsored projects/subprojects with substantial activity and progress and sponsor feedback (Appendix A) were measured by use of a Project Performance Measurement Questionnaire (Appendix B). Responses to all 32 questionnaires were obtained with the following results (The graph associated with each measure shows the performance measure over the last 26 fiscal years):

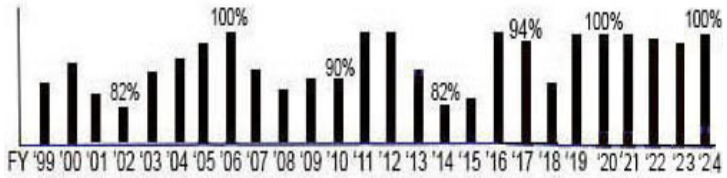
Measure 1. Overall, Work Met Expectations

Percent of FY2024 Program Sponsored Projects/Subprojects where sponsors reported that overall work met their expectations (agree or strongly agree) (32 out of 32 responses) ….. 100%



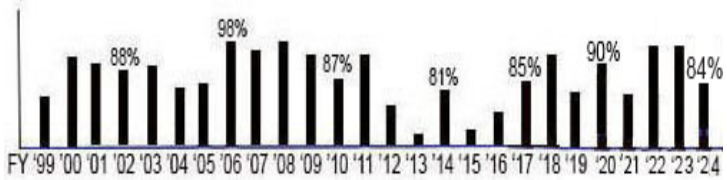
Measure 2. Established Major Deadlines Met

Percent of FY2024 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met (32 out of 32 responses) ……..…..…..….. 100%



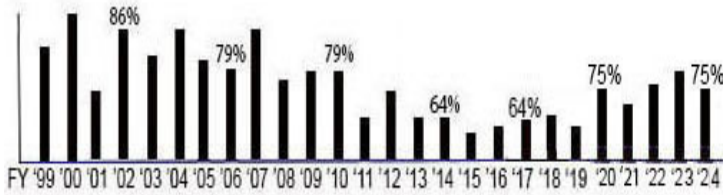
Measure 3a. At Least One Improved Method, Developed Technique, Solution, or New Insight

Percent of FY2024 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight (27 out of 32 responses) … 84%



Measure 3b. Plans for Implementation

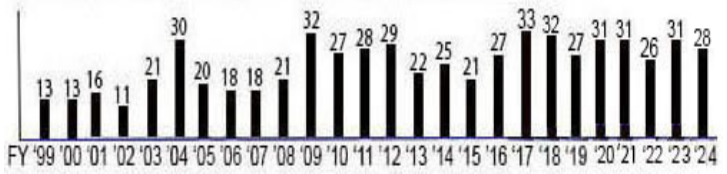
Of these FY2024 Program Sponsored Projects/Subprojects reporting at least one improved method, technique developed, solution, or new insight, the percent with plans for implementation (24 out of 32 responses) …….. 75%



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

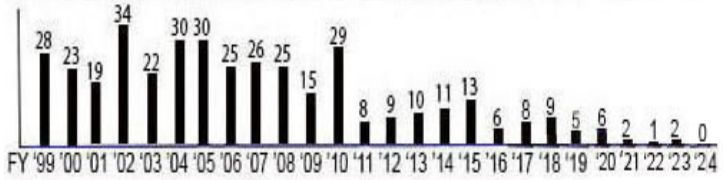
Measure 4. Journal Articles (Peer-Reviewed), Publications

Number of peer reviewed journal publications documenting research that appeared (16) or were accepted (12) in FY2024 …….. 28



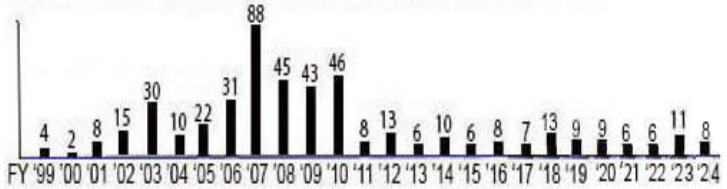
Measure 5. Proceedings, Publications

Number of proceedings publications documenting research that appeared in FY2024 ……..……. 0



Measure 6. Center Research Reports/Studies, Publications

Number of center research reports/studies publications documenting research that appeared in FY2024 ….. 8



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

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

TABLE OF CONTENTS

1. COLLABORATION

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

1.2 DEMOGRAPHIC FRAME 2030 (Decennial Project 5350K04)

1.3 RESPONSE DATA QUALITY 2030 (Decennial Project 5450K21)

1.4 RESPONSE PROCESSING PLANNING & SUPPORT (Decennial Project 5450K23)

1.5 REDISTRICTING DATA PROGRAM (Decennial Project 5550K02)

1.6 EVALUATIONS, EXPERIMENTS, & RESEARCH (Decennial Project 5650K01)

1.7 POST ENUMERATION SURVEY DESIGN & ESTIMATION (Decennial Project 5650K02)

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

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

Highlights: During FY 2024, staff worked on a major revision of a previously written and submitted paper. Staff completed the revision of the paper and has submitted the revised version to the journal for consideration. One major innovative use of administrative records was to enable classifying some addresses as occupied, vacant, or nonresidential when neither a self-response nor non-response follow-up response was available.

Staff: Mary Mulry (682-305-8809)

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

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

Highlights: During FY2024, the paper was published in the *Statistical Journal of the IOAS* (see publications section). One finding from the study indicates that both the mode of response and the amount of recall required of the respondent due to the length of time since April 1 affect the agreement rate between the census roster and the administrative records roster. This project is completed.

Staff: Mary Mulry (682-305-8809)

B. Supplementing and Supporting Nonresponse with Administrative Records

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

Highlights: During FY 2024, staff revised two draft memoranda documenting results of multinomial logistic regression models fit on 2010 administrative records (AR) housing unit (HU) data for non-response follow-up units (with 2010 Census Unedited File (CUF) HU size as

the dependent variable) and applied to the 2020 AR modeling HU data: one memorandum for the model with separate sub-models for different values of AR household size and one memorandum for the combined model (uses IRS 1040 household size as a predictor variable rather than having separate sub-models for different values of IRS 1040 household size). Staff revised the draft memorandum comparing outlier detection results for AR Modeling Closeout Deletes not assigned by AR by CUF Final Status. For AR Modeling Closeout Deletes not assigned by AR, the distributions of both ratio object score and individual modeling variables tend to be reasonably similar regardless of CUF Final Status.

Staff: Michael Ikeda (x31756)

C. 2020 Census Privacy Variance

Description: The Census Bureau is investigating the within run variance of the 2020 Census differential privacy (DP) algorithm. Specifically trying to identify the accuracy by which individual counts can be estimated given differing levels of released data. For a fixed privacy budget, this project treats true counts as unknowns and estimates them from the released differentially private data. This surmounts to understanding and solving what is known as a least absolute deviations regression. The ultimate objective is to explore via simulation the possibility of economizing on computation, to approximate the desired variance by actually computing simulated variances in slightly simpler problems with fewer marginal-total related observations.

Highlights: During FY 2024, progress was made in simulating various configurations of hypothetical Disclosure Avoidance System (DAS) runs. Multiple scenarios were crafted and executed as part of a factorial design, with each set of parameters being run multiple times. This thorough approach allowed for detailed analysis comparing the variance of full workloads versus partial workloads. Key insights emerged from modeling partial workload variances alongside covariate information, such as cell size, which enabled accurate predictions of full workload variance. This methodology aimed to provide computational savings from partial workload runs while offering full workload variance estimates.

Staff: James Livsey (x33517), Eric Slud

D. Experiment for Effectiveness of Bilingual Training

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

Highlights: During FY 2024, staff completed and released a Center for Statistical Research & Methodology *Research Report SSS2024-01* "A Multinomial Analysis of Bilingual Training and Nonresponse Follow-up Contact Rates in the 2020 Decennial Census." There is found not to be significant evidence to conclude that the training is associated negatively with response rate; i.e., training appears not to harm response rate. Findings are formulated in this doubly-negated way - rather than a more positive statement of training effectiveness - because of inherent limitations in the case data. Limitations include lack of exact indicators for which enumerators were bilingual and which households needed an interview in Spanish. Additionally, staff completed "2020 Census Spanish-Speaking Enumerator Training Experiment Report" for the Census Program for Evaluations and Experiments (CPEX) report series. This report is currently pending release and the project is completed.

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

E. Unit-Level Modeling of Master Address File Adds and Deletes

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

Highlights: During FY 2024, staff completed the writing of a final report on a tree-based predictive analysis of MAF ACS-universe status changes (both adds and deletes of MAF addresses in the ACS universe). This report was presented to the Decennial Research Objectives and Methods Working Group (DROM) during Q1. Revisions including some additional analyses

and rewriting, to satisfy reviewer comments have not yet been completed. The final report will be submitted in FY 2025. After the revision is re-submitted to DROM and also to the Disclosure Review Board, the report will be posted as a *CSRM Research Report*.

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

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

Highlights: During FY 2024, staff regularly participated in weekly meetings with the decennial coverage staff. Staff has reviewed and given feedback on plans for the 2030 Census Coverage Measurement program. One of the main questions still being discussed is whether field operations are needed to verify whether a census enumeration is correct or erroneous and if we can rely on administrative records for that determination. The administrative data comes from the Demographic Frames program which is an amalgamation of multiple administrative record sources. For national and state coverage estimates, dual system estimation and log-linear latent modeling (Van der Heijden) approaches are being explored. Small area methods are being considered for county and sub-county estimates.

Staff: Jerry Maples (x32873), Ryan Janicki

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

Highlights: During FY 2024, staff investigated the use of statistical models as well as Bayesian Additive Regression Trees for predicting block group reliability in the PL94-171 tabulations. Staff improved the tabulations and displays in an in-progress manuscript and considered new covariate values to increase the predictive accuracy of the models. Staff applied the developed models to the 2020 PL94-171 tabulations to estimate the block group reliability in the case where a swapping count is not available for directly computing the reliability. Staff observed that the estimated block group reliability in 2020 was comparable to the reliability observed in the 2010 data. (For related work on variance estimation, see the Subproject Q "Variance Estimation and Modeling for Privacy Protected Redistricting Data.")

Staff: Kyle Irimata (x36465), Tommy Wright

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

Highlights: During FY 2024, staff assisted with preparation of the 2020 Census Supplemental Demographic and Housing Characteristics (S-DHC) File for public release. These data include tabulations based on person and household characteristics jointly, at state and nation levels of geography. Here, the PHSafe disclosure avoidance algorithm - developed by Tumult Labs - added statistical noise to protect S-DHC tabulations. A Bayesian modeling framework was used to post-process raw PHSafe outputs and enforce constraints such as nonnegativity of counts. A manuscript describing the methodology was submitted to arXiv (https://arxiv.org/abs/2406.04448).

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

I.1. Imputation Modeling for Multivariate Categorical Characteristic Data in 2030 Census
Description: Nonresponse and administrative record enumeration in the Decennial Census led to item missingness for person and household characteristics. The 2020 Census used past census and administrative record data to directly assign characteristics when missing. For the 2030 Census, staff are researching

statistical imputation models for multiple categorical variables. The broad goal of this project is to study how to make better use of statistical modeling, in conjunction with administrative records, to enhance previously implemented procedures for characteristic imputation.

Highlights: During FY 2024, staff continued regular conversations with Research & Methodology Directorate and Decennial Statistical Studies Division colleagues to review techniques for multiple imputation with categorical characteristic information, discuss results from group quarters multiple imputation for categorical characteristic information, provide advice and suggestions for multiple chained equation methods in R, and learn from colleagues about multiple imputation with latent class models based on a large number of classes.

Staff: Darcy Steeg Morris (x33989), Joseph Kang, Joseph Schafer (R&M), Emanuel Ben-David

I.2. 2030 Characteristic Imputation Modeling Research for Nested Data with Structural Zeros

Description: The project is focused on advancing the methodology and implementation of edit and imputation for the 2030 Census. It aims to explore and develop methods to effectively impute missing data in nested data with structural zeros, even in situations where reported information is scarce or non-existent. The project will delve into statistical modeling for imputation in multi-level nested data, for example, data nested at the person level within households. The goal is to create a more robust imputation methodology, developed through a thorough and comprehensive process, with the potential to significantly enhance the analysis of future census data.

Highlights: During FY 2024, staff created a benchmark nested data set based on the 2012 Public Use Microdata Sample (PUMS). This data set included information at both the individual and household levels. We then simulated a missing data set using a missing at random mechanism as part of our first experiment to evaluate our edit and imputation strategies. We proposed four different strategies for editing and imputing the data. To assess the quality of the resulting edited and imputed data sets and compare them across different approaches, we generated 50 edited and imputed data sets for each strategy. We then compared the estimates produced by these data sets when estimating various population parameters, such as the proportion of households with at least two generations present and households with both partners being Hispanic. On average, the estimates obtained from the edited and imputed data sets across 50 replications closely matched those derived from the original sample data set, considered the gold standard data set.

Staff: Emanuel Ben-David (x37275), Tom Mule (R&M), Joseph Schafer (R&M)

J. Group Quarters Count Expectation Modeling to Ensure Data Quality

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

Highlights: During FY 2024, CSRM staff work on this project was suspended.

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

K. Mobile Questionnaire Assistance: Analysis and Simulation

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

Highlights: During FY 2024, staff developed a panel count model for tract-level response counts over intervals of time within a census operation. Uncertainty may be expressed with predictions through associated intervals. Daily self-response projections - produced by Census Bureau staff in earlier work - were considered as a component of the intensity function. Initial work on a regression form was considered to capture the relationship between MQA exposure and response rate. Characterizing this relationship is desired for potential application to model-assisted placement in future operations. CSRM staff began development of several R packages and a suite of associated documentation to facilitate the transition of research work to technical project staff.

Staff: Andrew Raim (x37894), Lisa Moore (DCMD)

L. Agreements for Advancing Record Linkage

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

University of Arkansas, Little Rock.

Highlights: During FY 2024, the universities have developed record linkage algorithms and methodology, published peer reviewed papers, and engaged in staff in bi-weekly meetings to include progress reports, presentations, and feedback from staff. We continue with this plan where research, meetings, and feedback are continuous.

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

M. Continuous Count Study

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

Highlights: During FY 2024, staff ran a simple imputation method on the 2020, 2021, and 2022 Demographic Frame Files and made the final output files (along with documentation) available to several Census Bureau researchers. The method replaces Demographic Frame File data with recoded Census Edited File (CEF) data for all persons who match to the CEF on PIK. Matching only on PIK (instead of on both PIK and MAFID) substantially increases the proportion of Demographic Frame File records that match to the CEF. For missing data remaining after this step, missing sex and age group are imputed with a simple previous person hot deck (with a geographic sort). Within-household missing race and Hispanic Origin are imputed using a previous person hot deck (or next person, if necessary) within the household. Whole-household missing race and Hispanic Origin are imputed using a previous person hot deck (with a geographic sort) where missing race is imputed using persons with the same Hispanic Origin and vice versa. Staff documented the results of the most recent imputation of the 2020, 2021, and 2022 Demographic Frame Files in three draft memoranda, one for each Demographic File. One key result is that using CEF data tends to decrease the proportion of White Hispanics and increase the proportion of Multiracial Hispanics. In addition, for Sex, Hispanic Origin, and Age category (adjusted for the change in reference date for the 2021 and 2022 comparisons), "assigned" and "as reported" CEF values are both nearly identical to the non-missing Demographic Frame File values. For Race, "assigned" and "as reported" CEF values have differences in similar directions ("assigned" tends to be less different) from the non-missing Demographic File values. Staff continued examining the use of IRS names data to impute sex on the Demographic Files. The generic SAS name parsing function is used to parse the relevant name field separately for each IRS form type tax year concatenated file (referred to as a "source"). Staff used three methods for assigning sex for each source: the first uses a SAS gender assignment function, the second assigns based on gender proportions for specific names at the national level, the third assigns based on gender proportions of specific names at the state level. Staff developed rules for combining the assignments from each source into an overall assignment. All three overall assignments performed well (the overall assignment based on the SAS gender function was not done for the 2022 Demographic File), but the assignment based on national level calculations did better than the assignment based on the SAS gender function especially when the two disagree. The assignment based on state level calculations performed similarly to the assignment based on national level calculations but assigned sex to a slightly lower proportion of records. Staff also defined a combined method which assigns if either the assignment based on state level or the assignment based on national level assigns unless the two assignments were contradictory. The combined method assigns sex to a slightly higher proportion of records than either assignment alone. Staff documented the results in seven draft memoranda. For the 2020 and 2021 Demographic Frame Files, one memorandum documented the results of the assignment based on the SAS gender function, one compared the results of the assignment based on the SAS gender function to the results of the assignment based on the national level calculations, and one compared the assignments based on the state and national level to the combined assignment. The comparison of the state, national, and combined assignments was also done for the 2022 Demographic Frame File. Staff examined the relationship between IRS status and age group after CEF replacement for the 2020, 2021, and 2022 Demographic Files. Staff divided the PIKs on the IRS files corresponding to each Demographic File into PIKs assumed to be dependents (IRS 1040 dependent PIKs) and PIKs assumed to be filers (IRS 1099 and 1099r PIKs and IRS 1040 primary and secondary PIKs). The IRS PIKs were matched to the relevant Demographic Frame File on PIK. The combination of IRS filer status and IRS dependent status strongly differentiates between children and adults among Demographic Frame File records with non-missing age group. Records that are both IRS filers and IRS dependents are not overwhelmingly either adults or children although they are usually in the boundary age groups (ages 10-17 and 18-29). Staff documented the results in three draft memoranda, one for each Demographic Frame File. Staff produced boxplots showing distributions of tract percent differences and absolute percent differences of eight sets of estimates (2020 and 2021 versions of three annual administrative

records (AR) population estimates and a calibrated AR DSE) from the 2020 Census tract populations by tract characteristics. Some of the boxplots used an arc tangent transformation that has minimal effect until the percent difference becomes very large. Modified versions of some of these boxplots were included in the report "Initial Results from the Continuous Count Study" for which Staff was included in the coauthors. The report and associated presentation slides were provided to the Census Scientific Advisory Committee and the National Advisory Committee as supporting material for their Spring 2024 meetings.

Staff: Michael Ikeda (x31756)

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

Highlights: During FY 2024, staff investigated whether coverage of the Demographic Frame could be improved by assigning more MAFIDs to IRS 1040 records through linkage with the 2020 Census on Protected Identification Key (PIK). It was found that there were substantial similarities in the address fields across databases, suggesting that additional MAFIDs may be assigned. Nationwide matches were placed into six categories based on the degree of address similarity and passed over to ERD for further investigation.

In another project, staff performed triple-system estimation of the count of Hispanic/non-Hispanic populations in 2020 using the American Community Survey, 2020 Census, and the Demographic frame. Staff presented their work at the 2024 Joint Statistical Meetings and will be publishing their work in the journal proceedings with the title: "Triple System Estimation of National Population Counts Through Log-linear and Latent Class Modeling."

Staff: Dan Weinberg (x38854), Tom Mule (R&M)

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

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

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

Highlights: During FY 2024, staff conducted a comprehensive evaluation of the utility of Numident in conjunction with the demographic frame for counting births by county. We focused on the count of 2020 births in Minnesota for proof of concept. Because the Numident has no information on the county of birth, we used the city of birth to determine its county. However, this approach can yield multiple counties for a small proportion of cities because these cities reside in more than one county, or various counties may have towns with the same common name. We undertook a thorough investigation to determine whether an imputation approach can impute the county of birth in cases where we cannot uniquely determine the county of birth. This approach is promising but requires imposing edit rules to restrict the choice of imputed counties to match the city of birth.

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

P. Cohort Component Domestic Migration Modeling (Continuous Count Study)
Description: The goal of this project is to utilize multiple data sources such as sample survey data, tax records, and other administrative data sources, to estimate the number of domestic migrants and the rate of domestic migration, as well as to provide uncertainty measures for the estimated counts and rates. Various modeling strategies will be explored to produce precise estimates of migration at low levels of geography (county, tract, block group) and for different cohorts (age, race, sex). This is part of a larger effort to develop a cohort component model for population which incorporates births, deaths, domestic migration, and international migration.

Highlights: During FY 2024, staff researched methods for estimating county level domestic migration by age race and sex using survey data and administrative records and other publicly available data sources. Regression models were developed by examining relationships between direct estimates using American Community

Survey data and other auxiliary data sources and the LASSO was used to select the most important predictors. Linear models and Poisson models were then fit to these data sets. Staff also researched extension to account for spatio-temporal dependencies.

Staff: Ryan Janicki (x37275), Serge Aleshin-Guendel, Tom Mule (R&M), Eric Jensen (POP), Esta Miller (POP)

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

Highlights: During FY 2024, staff provided software and technical support for Linux and Jupyter, staff wrote software to manage parallel runs of *BigMatch* for the Jupyter platform analogous to the 2010 tests runs. Staff supported software development in Management of Record Linkage in Linux. Staff also evaluated software packages SPLINK. By examining speed, accuracy and work requirements. *BigMatch* outperformed other Record Linkage softwares in timing and accuracy. Staff have archived software on Jupyter notebooks and github.

Staff: Ned Porter (x31798), Dan Weinberg

R. Variance Estimation and Modeling for Privacy-Protected Redistricting Data
Description: Starting with the 2020 Census, the Census Bureau implemented the TopDown Algorithm (TDA) to protect respondent confidentiality. The TDA incorporates differential privacy (DP), as well as post-processing steps to ensure that certain constraints and quality standards are met. Though the variability due to DP is publicly available, the variability due to post-processing is not as easily quantified. The goal of this project is to investigate methods for quantifying the overall variability due to the TDA in the PL94-171 redistricting data products using publicly available data.

Highlights: During FY 2024, staff investigated small area estimation approaches for estimating the variance of PL94-171 population counts which were privacy protected by the TDA. Staff developed variability estimates using a decile grouping by state for the total and racial subgroup counts of population at the block group level. Staff produced an extension to a joint Bayesian hierarchical model to produce estimates of the variability, while also producing model-based point estimates of the population counts using only publicly available data. Staff observed that the joint model produced similar or improved point estimates for racial subgroups as compared to the direct estimates obtained from the TDA. Staff showed that the joint model also produced variance estimates that were able to vary within a replication group, as compared to the estimates from decile groups.

Staff: Kyle Irimata (x36465)

1.8 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 6385K70)

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

Highlights: During FY 2024, staff submitted a manuscript that discusses methodology for generating custom ACS estimates from a continuous-time model for publication in a peer-reviewed journal. This manuscript has been revised subject to peer review and has been subsequently accepted for publication.

Staff: Patrick Joyce (x36793), Tucker McElroy (R&M)

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

Highlights: See reference in above *Description*.

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

C. Voting Rights Act (VRA) Section 203 Research Towards 2026 Determinations (also Decennial Project 6550J06)
Description: The *Voting Rights Act* of 1965 prohibits discrimination in voting. Section 203 of the *Voting Rights Act* mandates the Census Bureau to make estimates every five years relating to totals and proportions of citizenship, limited English proficiency and limited education among specified small subpopulations (voting-age persons in various race and ethnicity groups called Language Minority Groups [LMGs] for small areas such as counties or minor civil divisions MCDs); these estimated total counts and proportions are used to determine which of nearly 8,000 jurisdiction must provide voting materials in languages

in addition to English. Specifically, the Section 203 determinations result in the legally enforceable requirement that certain jurisdictions (e.g., states, counties, cities, etc.) 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 estimation models based on 5-year American Community Survey (ACS) data, the Decennial Census, and possibly administrative records. These models will be used to produce more accurate estimates in small areas for the 2026 determinations.

Highlights: During FY 2024, staff accomplished the following: (1) implementation of the Bayesian model using the Polya-Gamma distribution with its R package BayesLogit for enhanced computational efficiency and stability compared with the 2021 VRA models, (2) presentation of improved diagnostic results to the steering committee and to the statistical audiences at the Royal Statistical Society in the UK.

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

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

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

Highlights: During FY 2024, staff continued regular conversations with colleagues in the Decennial Statistical Studies Division, Center for Economic Studies, and Social, Economic, and Housing Statistics Division to understand post-collection processes including calibration to population estimates, variance estimation, and geographic levels. Staff continued work with a research dataset on response status for and administrative records for 2018-2022 ACS sampled units. With this dataset, staff developed and compared logistic regression, lasso, classification trees, random forest and generalized boosting models for modeling response propensity. Staff continued to refine the models based on ACS process and theoretical considerations understood through regular conversations with ACS experts. Staff continually update analysis to reflect developing best practices and current developments for machine learning algorithms with big survey datasets, e.g., classification trees with probability based stopping rules (*ctree* in R), random forest tuning without variable selection, and a faster boosting method shortcut fitting method (*lightgbm* in R). These model developments led to response probability estimates that are used in an alternative nonresponse adjustment approach that (1) adjusts weights with the inverse of the response probability estimate to balance respondent and nonrespondent characteristics and then (2) calibrates to benchmarks. Staff continued evaluating metrics to use for model comparison to include practical considerations as well as traditional checks such as comparison of weighted means of respondents to the entire sample for a large set of characteristics. Staff continued documenting methods and results for an internal report. Staff also began developing geographic visualizations of nonresponse model comparisons to illuminate state-level variation in lack of fit.

Based on propensity model evaluation metrics, logistic regression and GBM were further studied by implementing the full ACS data processing with these two options as alternative weight methodologies for all states and all years. Staff developed metrics and interactive plots to assess differences in ACS outcome estimates and variance estimation for the various weighting methodologies: production, entropy balance weighting, and inverse probability weighting (IPW) (with logistic and GBM response model). Staff assessed ACS outcome estimates using short time series modeling to assess (in)consistency in the different methods over time. This work was presented at the Small Area Estimation conference in Lima, Peru.

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

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 FY 2024, staff refined and further validated the simulation code, extended a theoretical formula for relative SDR bias in the context of Stratified SRS samples, and did additional writing on a draft manuscript. The new written material is an exposition on the (rather modest) typical biases due to SDR in small domains, also elaborating the special conditions involving alternating behavior in neighboring strata under which more serious biases in SDR variance estimation in small domains can arise. The manuscript is almost complete and will be reviewed and submitted in FY 2025.

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

B. Nonresponse Adjustments for High Frequency, Low Response Surveys

Description: Demand for timely data on current topics has motivated the development of high frequency web-based surveys, specifically the Household Pulse Survey. Such surveys can exhibit low response rates and complex missing data that may leave them particularly susceptible to nonresponse bias. This research assesses nonresponse bias through external benchmarks and internal administrative data, and studies traditional and novel application of nonresponse adjustment (e.g., calibration, inverse probability weighting) for low response surveys.

Highlights: During FY 2024, Center for Statistical Research & Methodology staff worked with a team of Demographic Statistical Methods Division staff on nonresponse adjustment for the Household Pulse Survey (HPS). Staff got familiar with technical details of HPS processing including imputation and calibration; reviewed questionnaires and production nonresponse bias reports. Staff assessed estimates for HPS outcomes as compared to available ACS estimates for select overlapping household characteristics (e.g., income, household structure, employment) to evaluate further calibration adjustments guided by assessing tradeoffs between possible estimate bias and imputation bias. Staff formulated a research plan for studying and assessing smoothing methods for extreme weights using a variety of modeling techniques including machine learning algorithms. And staff began applying geographically-varying visualizations of nonresponse patterns and outcome estimation bias as applied to potential nonresponse adjustment procedures for the HPS. As the HPS is absorbed into the Census Household Panel, staff is working to transfer lessons learned to the new format, reconsidering variables and domains for calibration as well as model-based techniques.

Staff: Darcy Steeg Morris (x33989), Sixia Chen (The University of Oklahoma), John Chestnut (DSMD), Mark Bauder (DSMD)

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

A. Describing Current Population Survey Field Representatives Reported Beliefs Regarding What Is Effective in Gaining Cooperation

Description: During August-September 2007, five hundred and forty (540) Field Representatives (FR) were randomly selected from the Census Bureau Regional Offices who worked on the Current Population Survey (CPS) to respond to a CPS Field Experiences Questionnaire as part of a new study. The study, conducted by researchers in the Center for Statistical Research and Methodology (formerly Statistical Research Division) was conducted with the broad purpose of aiming to improve the methods and procedures used for interviewer-administered questionnaires. Specifically, the questionnaire asked questions grouped in three sections: Section I: asked about FR's background and experiences working with the CPS (5 questions); Section II: asked questions about the FR's beliefs about various practices/techniques effectiveness in gaining cooperation from CPS selected households (60 questions); and SECTION III: asked FR how often he/she used various practices/techniques (10 questions). Each FR in this study received one hour pay to complete and return the questionnaires. Five hundred and twelve (512) completed questionnaires were returned. A database of responses was developed, and some limited analyses were started, but not completed. The purpose of this project is to report tables and descriptions of what is in the database with the hope of providing clues for further focused research and experiments. We will also attempt to investigate relations between FR beliefs and FR CPS interview completion rates.

Highlights: During FY 2024, staff completed a draft report which is near ready for review: "Analysis of Field Representatives Beliefs on Effectiveness of Practices in Gaining Cooperation from Households for the Current Population Survey." The draft contains the following sections: Abstract, Introduction, FR Beliefs about Effectiveness of Practices, How often FRs Report Using Certain Practices, and Relation between a Practice & Completion Rate. Among the draft report's findings are:

(1) 85% of FRs reported the belief that the practice "smiling and having a friendly facial expression" has "extremely positive effects" in gaining cooperation for a CPS interview [next three highest beliefs were in the practices- "having good appearance (e.g., be neat, well-groomed, look professional)", "showing my ID badge as soon as anyone opens the door", and "stressing the confidentiality of respondent's reported data"]; (2) 90.4% reported using the following practice either "most" or "all" of the time-- "dressing professionally for personal visits;" and (3) the top four of the eight practices that showed a statistically significant relation with completion rates (significance level 0.10) were: "sending thank you notes with wording provided by the Regional Office," "carrying personal items/gifts for respondents (s) to help gain trust/cooperation (e.g., dog treats, small gifts for children, flowers, candy, etc.), "contacting a respondent at his/her place of work to complete an interview," and "keeping and continuing to work on a reluctant household instead of requesting assignment of household to another FR."

Staff: Tommy Wright (x31702), Joseph Engmark, Thomas Petkunas

B. Using Machine Learning for Improving Nonresponse Adjustment in the Current Population Survey

Description: The response rates to the Current Population Survey (CPS) have declined in recent years. This decline has raised concerns about potential bias in key population statistics due to nonresponse. An effective way to address this is by using administrative data to adjust the weights for nonresponse while keeping the calibration of population estimates unchanged. This involves linking the administrative data to both responding and non-responding households in sample surveys. Once linked, we can use the linked data to adjust the weights for respondents to account for differential nonresponse rates among different subpopulations. In this project, we propose two main aspects. First, we aim to enhance nonresponse adjustment by using more advanced machine learning models. Second, we aim to address potential errors in the linkage process, which can impact the performance of models used for nonresponse adjustments.

Highlights: During FY 2024, staff reviewed the most recent research proposals for addressing nonresponse units in the current population survey. Staff identified and obtained access to several administrative datasets that could assist in better weighting the nonresponses. Additionally, staff recruited new members and scheduled biweekly meetings to ensure more thorough involvement, discussion, and progress.

Staff: Emanuel Ben-David (x37275), Tim Trudell (DSMD), James Ross Foy Ohagan (DSMD), Jonathan Eggleston (CES)

C. Data Integration

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

Highlights: During FY 2024, staff wrote data cleaning software for curbstoning research. The members of the Improving Quality for Enumeration team provided feedback on such research. For the reidentification, staff is developing software to standardize and clean data from Public Use files and Administrative List files. These records need to be standardized to make it easier to link records. Separate data sources define data for example would be income could be a range or a specific number. Other data can be categorical with different values for the same variables or just differently defined metadata. Those data are being standardized for possible linkage to test vulnerable data. The software is archived on github.

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

1.11 SOCIAL, ECONOMIC, & HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165024)

A. Research for Small Area Income and Poverty Estimates (SAIPE)

Description: The purpose of this research is to develop, in collaboration with the Small Area Estimates Branch in the Social, Economic, and Housing Statistics Division (SEHSD), methods to produce “reliable” income and poverty estimates for small geographic areas and/or small demographic domains (e.g., poor children age 5-17 for counties). The methods should also produce realistic measures of the accuracy of the estimates (standard errors). The investigation will include assessment of the value of various auxiliary data (from administrative records or sample surveys) in producing the desired estimates. Also included would be an evaluation of the techniques developed, along with documentation of the methodology.

Highlights: During FY 2024, staff (along with their SEHSD counterparts) have been assisting the Data Integration subteam of the Continuous Count Study project. The modeling framework comes from the small area estimation share model developed for school district poverty estimates.

Staff have been working on how to construct share models that contain both structured and unstructured dimensions for the domains of estimation. Specifically, the domains of estimation are tract by demographic subgroups. The demographic groups (8 race groups by 2 age groups) are the same across all tracts [structured], but the number of tracts per county vary and their order is not relevant for estimation [unstructured]. The Dirichlet-Multinomial share model is based on unstructured domains. Results from this work are expected to inform model choices for school district estimation.

Staff met with colleagues in SEHSD to start developing an evaluation plan for the county and school district models.

Formal evaluations of new statistical models for the county and school district level poverty estimates are being planned for FY 2025.

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

B. Assessing Constant Parameters across Areas in the SAIPE Models

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

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

Highlights: During FY 2024, staff derived a new method to test whether model parameters vary from the constant parameter model using geographically weighted small area models. The method is based on setting up a series of estimating equations, one for each location or area of interest, and using sandwich variance estimators to obtain a test statistic. Even though the number of parameters is larger than the number of small area domains, the special block diagonal structure gives a form that allows consistent estimates of the covariance matrix of the parameters. Staff gave a presentation on geographically weight small area models at the Small Area Conference in Lima, Peru in June using county level poverty estimates as the data example (public use version of the data).

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.

Highlights: During FY 2024, no significant progress was made on this project.

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

1.12 GENERAL ECONOMIC STATISTICAL SUPPORT (Economic Project 1183X01)

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

A. Use of Big Data for Retail Sales Estimates

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

Highlights: During FY 2024, staff worked with colleagues in the Economic Statistical Methods Division to develop hierarchical Bayesian mass imputation methodology for estimating state-level retail sales based on data from a third-party aggregator. This revised paper is now published in the *Journal of Survey Statistics and Methodology*. Staff presented this work -- and the related work using a different methodology and different data source -- at the 2024 International Conference on Establishment Statistics and in an invited talk at the 2024 Joint Statistical Meetings.

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

B. Seasonal Adjustment Support

Description: This is an amalgamation of projects whose composition varies from year to year but always includes

maintenance of the seasonal adjustment software used by the Economic Directorate.

Highlights: During FY 2024, staff provided assistance with helpline requests from various organizations, including the Pennsylvania Department of Labor & Industry, University of Basel, Accenture, Reserve Bank of India, Estima, Union Investment Privatfonds, Bank of Canada, Australian Bureau of Statistics, and Statistics New Zealand. Throughout the year, support was consistently extended to all open requests, ensuring ongoing collaboration and guidance.

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

C. Seasonal Adjustment Software Development and Evaluation

Description: The goal of this project is a multi-platform computer program for seasonal adjustment, trend estimation, and calendar effect estimation that goes beyond the adjustment capabilities of the Census X-11 and Statistics Canada X-11-ARIMA programs, and provides more effective diagnostics. The goals for FY 2024 include: continuing to develop a version of the X-13ARIMA-SEATS program with accessible output and updated source code so that, when appropriate, the Economic Directorate can produce SEATS adjustments; and incorporating further improvements to the X-13ARIMA-SEATS user interface, output and documentation. In coordination and collaboration with the Time Series and Related Methods Staff of the Economic Statistical Methods Division (ESMD), staff will provide internal and/or external training in the use of X-13ARIMA-SEATS and the associated programs, such as X-13-Graph, when appropriate. Additionally, development efforts are focusing on the future software products that advance beyond current capabilities of X-13ARIMA-SEATS. This new product aims to handling sampling error, treatment of missing values, and multivariate analysis. This development is a joint effort with staff from the Center for Optimization & Data Science and the Economic Statistical Methods Division.

Highlights: During FY 2024, staff implemented a series of key updates and fixes for Build 61 of X-13ARIMA-SEATS. These enhancements included new features such as displaying error messages and halting the program under specific conditions. For instance, the program now stops if a specification file requires modeling (e.g., regression, check, estimate, forecast, outlier, or seats specs) but provides no modeling specs (such as arima, automdl, or pickmdl), or if an arima spec is included without a model argument. Additionally, the length of the Henderson filter used in the D7 table (d7trendma) was added to the .udg file.

Several defects from previous versions were also fixed. These fixes addressed issues such as:

- Missing data and values in the .rts file when the estimate spec includes save = rts (ASCII version).
- An issue when running with the -s command and the estimate spec does not include save = rts.
- Incorrect key values for the significance test for user-defined regressors in AICC testing.
- Errors in the regression matrix when the chi2test option was used and user-defined holiday regressors were removed.
- Incorrect program output for AICC testing when a p-value for AIC testing was specified.
- An error message related to the start date of the original series when the x11regression spec was included.
- Errors preventing history analysis when the history spec included print = none.

The release also addressed minor text typos. Build 61 was made publicly available at https://www.census.gov/data/software/x13as.X-13ARIMA-SEATS.html.

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

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

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

behavior under long range dependence and/or extreme values.

Highlights: During FY 2024, advancements were made in modeling weekly data using a non-integer periodicity differencing operator. This enhancement involved specifying a unit root structure for time series with fractional periodicities, which converges to standard seasonal differencing operators with integer periodicities. The weekly unit root operator was further examined to clarify the precise location of each unit root along the unit circle, allowing users to easily visualize the functional form applied to their time series. All related code has been implemented in Ecce Signum (sigex), our custom structural component model codebase, available at www.github.com/tuckermcelroy/sigex.

Additionally, in efforts to avoid residual seasonality in hierarchically adjusted time series, particularly focusing on U.S. GDP and its sub-components, we completed a comprehensive research report. This report is publicly available for reference at https://www2.census.gov/library/working-papers/2024/adrm/RRS2024-02.pdf.

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

E. Supporting Documentation and Software for Seasonal Adjustment

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

Highlights: During FY 2024, we implemented changes and corrections to our X-13ARIMA-SEATS documentation. This included addressing minor errors such as typos and formatting inconsistencies and making substantial updates to clarify critical functionalities. Chapter 4 was enriched with a discussion on the effects of inserting a sequence of outliers at the end of a series. We also revised the descriptions of the appendbcst and appendfcst arguments, clarifying their application across composite, seats, series, and x11 contexts. A detailed analysis of the maxlead argument in the series specification was introduced, and an important correction was made to the flags associated with the excludefcst argument in the x11 specification, ensuring they correctly reflect no/yes/no settings instead of yes/no/yes. To support these updates, additional references were integrated to enhance the comprehensiveness of the documentation.

All of these changes were published as part of Build 61 of X-13ARIMA-SEATS, available at https://www.census.gov/data/software/x13as.X-13ARIMA-SEATS.html.

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

F. Exploring New Seasonal Adjustment and Signal Extraction Methods

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

Highlights: During FY 2024, we completed a simulation study that demonstrated the potential of multivariate signal extraction to outperform univariate methods in some cases, while showing comparable results in others. Alongside this, we conducted an empirical analysis of the M3 aggregates for indirect adjustment. This analysis involved reviewing revision histories, comparing the results with X-11 direct adjustment, diagnosing residual seasonality, and evaluating data smoothness measures. These findings were presented at a CSRM Seminar, and the seasonal adjustment practitioner's workshop.

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

G. Production and Dissemination of Economic Indicators

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

Highlights: During FY 2024, there was no significant progress.

Staff: Adam Hall (x32936)

H. Small Area Estimation for the Annual Integrated Economic Survey

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

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

Highlights: During FY 2024, staff expanded the research data set based on the 2017 Economic Census to include sales, employment, annual payroll, and first quarter payroll for all industries defined by 3-digit NAICS code groups. Staff have transitioned to using the County Business Pattern (CBP) dataset in place of the Business Register for predictor variables, which they have found to not have an issue with implausible outliers. Staff members have also transitioned the sample survey variance estimation used in the research data set to match the variance estimation being used for the AIES. Staff are creating a new research data set for all industries defined by 4-digit NAICS code groups, to explore the feasibility of small area estimation modeling at a finer industry scale.

Staff explored different covariate models for incorporating the CBP data into Fay-Herriot style small area models for the research data set. As the AIES sample survey design uses certainty strata, where some establishments are always included in the sample, staff also investigated the value of modeling outcomes separately based on certainty strata inclusion. Staff found there were convergence issues with using the Stan probabilistic programming language for model fitting and transitioned to using a rejection sampler for model fitting. An R package which implements the rejection sampler, rejectFH, has been developed, and is planned to be submitted to CRAN in FY2025. Staff are now conducting a prior sensitivity analysis for the chosen small area models.

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

1.14 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, Joseph Kang, Eric Slud, Kelly Taylor

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

Highlights: During FY 2024, the IRE Technical Team supported researchers at the Census Bureau, the Bureau of Economic Analysis, and at Federal Statistical Research Data Centers (FSRDCs) across the nation as they worked on over 1400 projects. We updated and maintained several statistical software packages in the Integrated Research Environment (IRE) – Stata, StatTransfer, and SAS, and added many user contributed software components for R, Anaconda python, and Stata. We acquired a new Matlab component – Matlab Compiler – which will be made available during the next update cycle of Matlab currently planned for November 2024. At the same time the team continued working on documenting the implementation of security controls in support of the security reaccreditation of IRE and the Cloud Research Environment (CRE). We also introduced the base version of Julia to IRE and continue to research strategies to increase its functionality by adding components while maintaining our current "isolated" security posture.

Staff: Chad Russell (x33215)

1.15 NATIONAL CANCER INSTITUTE (Census Bureau Project TBA)

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

Highlights: During FY 2024, staff performed Bayesian hierarchical modeling to produce county-level direct survey-based estimates for eight tobacco smoking outcomes. County-level designed-based estimates for these tobacco outcomes were produced for 3,134 counties across the United States.

Additionally, model-based estimates for population coverages for these eight tobacco smoking outcomes were produced for 3,134 counties using 2018-2019 Tobacco Use Supplement to Current Population (TUS-CPS) files. Bayesian Hierarchical modeling through a Markov Chain Monte Carlo (MCMC) simulation study was used to produce the final model-based estimates for states and county-level estimates for these eight tobacco outcomes.

The MCMC was conducted to compare four different small area estimation models for producing states and county level estimates. The four small area estimation models considered include: (a) Area-level probability small area model; (b) Arcsine-scale area-level small area model; (c) Area-level binomial-logit small area model; and (d) Beta-logistic model with unknown sampling variance.

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

2. RESEARCH

2.1 GENERAL RESEARCH AND SUPPORT (Census Bureau Project 0331000)

Missing Data & Observational Data Modeling

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

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

Potential Applications:
Research on missing data leads to improved overall data quality and estimate accuracy for any census or sample survey with a substantial frequency of missing data. It also leads to methods to adjust the variance to reflect the additional uncertainty created by the missing data. Given the ever-rising cost of conducting censuses and sample surveys, imputation for nonresponse and statistical modeling for using administrative records or alternative source data is important to supplement actual data collection in situations where collection is prohibitively expensive in Decennial, Economic and Demographic areas.

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

Highlights: During FY 2024, staff worked towards a deeper understanding of traditional missing data methodologies such as imputation and nonresponse weighting, with the purpose of re-thinking these methods in light of large-scale and biased missingness from decreasing response rates, data collection interruptions and survey design. To this end, staff is studying historic and recent literature on inverse probability weighting (IPW) and calibration techniques for nonresponse.

Staff continues hands-on learning of such methods through research motivated by studying the ACS experimental weighting calibration procedure for 2020 ACS data products and working groups sharing ideas for alternate methodologies. Motivated by this application, staff has been researching alternative weighting methodologies and implemented a study of a two-stage (inverse probability followed by calibration) approach for adjusting for dynamic and large magnitude unit missing data. This study developed knowledge in response propensity modeling with machine learning techniques, as well as study of how machine learning techniques can be incorporated into traditional survey processing. As a byproduct of using machine learning nonresponse modeling, staff is studying the proper use of survey weights in the algorithms and adaptations to evaluation that are tailored to the survey context. During Q3 of 2024, this work further involves research on variance estimation techniques with traditional models and machine learning techniques including developing appropriate metrics to evaluate and compare potential bias in calculated variances.

Staff is also building knowledge on modeling to jointly edit and impute for multivariate categorical variables. Motivated by the related project for 2030 Census characteristic imputation, staff continued learning

literature on simultaneous edit and imputation via Bayesian hierarchical models and began learning about large class latent class models for multiple imputation with the goal implementation of a model-based multiple imputation study on 2020 Census data.

Staff also continues to build knowledge and experience with traditional calibration methods but applied to nontraditional data that exhibits high nonresponse and a modern sampling strategy. Staff has assessed estimates for the Household Pulse Survey as compared to comparable ACS estimates to develop a strategy for further nonresponse adjustments guided by assessing tradeoffs between possible estimate bias and imputation bias. As the Household Pulse Survey is absorbed into the Census Household Panel, staff is working to transfer lessons learned to the new format, reconsidering variables and domains for calibration as well as model-based techniques.

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

B. Imputation and Modeling Using Observational/Alternative Data Sources

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

Highlights: During FY 2024, staff worked with Economic Statistical Methods Division (ESMD) staff to finalize a research paper – that has now been published in the *Journal of Survey Statistics and Methodology* -- describing the hierarchical Bayesian mass imputation model methodology for estimating state-level retail sales based on data from a third-party aggregator. An imputation model is built using the third-party data and applied to obtain imputations for all establishments in the survey frame. The imputed dataset is then used as input for the Monthly State Retail Sales (MSRS) – a more geographically granular and timely estimate than the produced Monthly Retail Trade Survey (MRTS). The purpose of the paper is to illustrate the usefulness of Bayesian multiple imputation hierarchical models (and ease of fitting with off-the-shelf software) for official estimates about the economy using third party data. This work, along with related work, was presented at the 2024 International Conference on Establishment Statistics and the 2024 Joint Statistical Meetings.

Staff is also assessing and studying the use of administrative record information in traditional imputation and nonresponse weighting methodologies. The projects described in Part I (Collaboration) introduce the novelty and availability of administrative record data to serve as both predictors in imputation and response propensity models, as well as to serve as benchmarks in calibration approaches. As part of those research projects, staff is interested in developing procedures for proper use and proper uncertainty quantification when using alternative data sources in missing data models. For example, staff is assessing estimation of variance for ACS estimates using machine learning IPW nonresponse adjustments with replicate weights at the modeling level as opposed to replication weighting factors at an aggregated level.

Staff continued developing methods regarding the use of alternate data in low response surveys with potentially large scale nonresponse bias specifically in light of nontraditional sampling methodologies (e.g., the Household Pulse Survey). For sample surveys with a less than perfectly defined sampling frame and a large-scale quantity of nonrespondents, the use of administrative records in nonresponse modeling is not as seamlessly applied as in previous work, thus staff has been studying the use of administrative record information as benchmarks for calibration adjustment and a source of comparison across respondent and nonrespondent populations rather than incorporating directly into nonresponse modeling.

Staff: Darcy Steeg Morris (x33989), Joseph Kang, Isaac Dompreh, Yves Thibaudeau, Patrick Joyce

Record Linkage & Machine Learning

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

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

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

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

Highlights: During FY 2024, the staff continued developing efficient methodologies for adjusting the statistical analyses for bias or inconsistency introduced by linkage error. Specifically, we designed a few methods for adjusting modern predictive ensemble models, such as bagging and random forests, for the presence of mismatch errors in linked data sets. Our proposed approach, based on mixture modeling, is not only efficient but also adaptable, accommodating various predictive ensemble modeling techniques in a unified fashion. We evaluated the performance of our proposed methodology with simulations implemented in R.

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

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

Highlights: During FY 2024, staff expanded their methodology to allow a larger number of matching fields as may be encountered in real-world settings. They presented their work with a lightning talk/poster at the 2024 Statistics and Data Science Symposium entitled "Using Orthogonalized Design Matrices to Improve Estimation of Record Linkage Parameters."

Staff: Daniel Weinberg (x38854), Yves Thibaudeau

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

Highlights: During FY 2024, staff published a paper which proposes new convergence diagnostics that are more appropriate and practical for illustrating a Markov chain Monte Carlo sampler that fails to converge. Furthermore, staff illustrated this using methodology and experimental results and provided open source software.

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

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

Highlights: During FY 2024, staff reviewed and proposed blocking methods and distance metrics for record linkage. The methodology was explored on both real and simulated data sets, and a paper was accepted for publication based upon the results.

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

(University of Connecticut), Ahmed Soliman (University of Connecticut)

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

Highlights: During FY 2024, staff, in collaboration with a Ph.D. candidate, proposed a record linkage algorithm, which links roughly half a million records in minutes.

During FY 2024, staff developed two variational approximations for bipartite record linkage, considering a mean-variational approximation and a stochastic approximation. We have applied our proposed methodology to both simulation studies and real applications, illustrating comparable accuracy to MCMC based methods. During Q4, we revised our manuscript for publication and submitted it along with working on our open source software package and documentation.

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

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

Highlights: During FY 2024, the staff finished revising the manuscript submitted for this research, which has been accepted for publication in *The Annals of Applied Statistics.*

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

Sampling Estimation & Survey Inference

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

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

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

Research Problems:
• How can methods making additional use of administrative records, such as model-assisted and balanced sampling, be used to increase the efficiency of household surveys?
• Can non-traditional design methods such as adaptive sampling be used to improve estimation for rare characteristics and populations?
• How can time series and spatial methods be used to improve ACS estimates or explain patterns in the data?
• Can generalized weighting methods be implemented

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

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

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

Highlights: During FY 2024, staff added 1-Year American Community Survey (ACS) data for 2022 to the comparisons of each state with the other states and for the estimated overall rankings of the states and a joint confidence region for 88+ different American ACS topics. The 2 updated visualizations now provide ACS data for the years 2018, 2019, 2021, and 2022 as parts of The Ranking Project on the Center for Statistical Research & Methodology's Internet site under "Statistical Research." See the three links The Ranking Project; Comparisons of A State with Each Other State; and Estimated Rankings of All States.

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

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

Sample Allocation
Highlights: During FY 2024, staff followed up on published work by Klein, Wright, and Wieczorek (2020) who presented a simple novel measure of uncertainty for an estimated ranking of K populations using a joint confidence region which tends to reveal more uncertainty for the middle of an estimated ranking than for the extremes. Staff published a *CSRM Research Report* (*RRS2024-01*) that explores and provides a framework which permits some control over the amount of uncertainty in various portions of the estimated ranking

with an optimal allocation of sample among the K populations. The *Research Report* also includes a definition of "tightness" that quantifies the amount of uncertainty in a joint confidence region.

Staff: Tommy Wright (x31702)

Apportionment

Highlights: During FY 2024, staff (1) worked on understanding the flaw in Hill's procedure for implementing Hill's Method of Alternate Ratios which eventually led to the Method of Equal Proportions, the current method of apportioning the 435 seats in the U.S. House of Representatives; (2) investigated a framework for determining an optimal size of the U.S. House of Representatives; and (3) investigated a mathematical framework for generalizing the Method of Equal of Proportions.

Staff: Tommy Wright (x31702)

C. A Joint Confidence Region for a Ranking Based on Differences

Description: Klein, Wright, and Wieczorek (2020) presents a simple novel measure of uncertainty for an estimated ranking by constructing a joint confidence region using overlapping intervals of separate population parameters for the unknown true ranking of K populations, k = 1, 2,…, K. Using the definition of "tightness" introduced in Wright (2024), it is desired to investigate what can be done to make joint confidence regions tighter.

Highlights: During FY 2024, staff conducted research and published a *CSRM Research Report (RRS2024-04)* that considers the same problem as Klein, Wright, and Wieczorek (2020), hereafter KWW, and Wright (2024) which present a simple novel measure of uncertainty for an estimated ranking by constructing a joint confidence region (INDI) using overlapping intervals of "individual" population parameters for the (unknown) true ranking of K populations, for k = 1, 2,…, K. In the new method, we consider the same problem but construct a new joint confidence region (DIFF) using intervals for "differences" of two population (k and k*) parameters. With this new approach and for each k and k*, we ask how many such intervals of differences with other populations overlap the number 0. We refer to this new joint confidence region by DIFF for differences. In addition to earlier research, we proved a condition under which DIFF shows no greater uncertainty than the uncertainty of the INDI joint confidence region.

Staff: Tommy Wright (x31702)

Small Area Estimation

Motivation: Small area estimation is important in light of a continual demand by data users for finer geographic detail of published statistics. Traditional demographic sample surveys designed for national estimates do not provide large enough samples to produce reliable direct estimates for small areas such as counties and even most states. The use of valid statistical models can provide small area estimates with greater precision, however bias due to an incorrect model or failure to account for informative sampling can result. Methods will be investigated to provide estimates for geographic areas or subpopulations when sample sizes from these domains are inadequate.

Research Problems:

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

Potential Applications:

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

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

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

the use of asymptotic expressions with a nonparametric bootstrap method to approximate the MSEs.

Highlights: During FY 2024, staff is reorganizing the initial draft of the work. Currently, they are working to rigorously justify the informal asymptotic expressions of the MSE of the EBLUP and the validity of the bootstrap estimator of the MSE. As part of the reorganization of the earlier draft, a technical supplement is currently under preparation.

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

B. Bayesian Hierarchical Spatial Models for Small Area Estimation

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

Highlights: During FY 2024, a staff member generalized his paper with Dr. Chung published in 2022 in the *Survey Methodology* journal titled "Bayesian Spatial Models for Estimating Means of Sampled and Unsampled Small Areas." Two authors collaborated with Dr. Li and Dr. Okech to prepare another manuscript. In the new manuscript, they developed noninformative hierarchical Bayesian prediction of small area means based on certain summary data. Summary data are not necessarily for individual states, but they are for various predefined partitions of the set of states in the study. In fact, the summary data may be direct estimates for a group of sub-areas (counties within a state) where the interest is in production of small area estimates for the sub-areas. This setup provides a generalization of the sampling part of the model in the above-mentioned *Survey Methodology* paper. For the present general setup none of the small areas (or sub-areas) may have any direct estimates. This work leverages use of available sub-area (county, for example) level covariates and state- or higher area-level direct estimates to develop reliable sub-state level characteristics of interest. This work considers various spatial models as linking models for the sub-state means to implement noninformative hierarchical Bayesian estimation. A real utility of this work is that it computes sub-state level estimates with an estimated measure of uncertainty (posterior variances or credible intervals) when no direct estimates at the sub-state level exist. A simulation study also shows the usefulness of this work. A manuscript based on this work has been accepted for publication in a special issue of *Statistics and Applications* in memory of C.R. Rao.

This work may have some potential to the Annual Integrated Economic Survey study on which some of staff from the Center for Statistical Research & Methodology are collaborating.

Staff: Gauri Datta (x33426), Ryan Janicki, Jerry Maples, Hee Cheol Chung (UNC Charlotte), Jiacheng Li (Wells Fargo), David Okech (University of Georgia)

C. Exploration of Small Area Estimation via Compromise Regression Weights

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

Highlights: This project is completed.

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

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

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

Highlights: During FY 2024, staff collaborated on a manuscript detailing a novel Bayesian approach to construction of joint credible set of ranks of small area means. The method is based on some suitable joint credible sets of small area means. Using such a credible set of the means, staff created joint posterior distributions

of ranks of all small areas under consideration. These distributions are called the empirical credible distributions of the ranks. The empirical credible solution compares favorably with the available frequentist solution based on joint confidence set of ranks. Datta collaborated with two external members to prepare a manuscript for publication. A revised version of this manuscript is currently being reviewed by a journal.

Moreover, two staff members (Datta and Maples) are collaborating on multiple projects based on the topic credible distributions of ranks of entities. With a few other external collaborators, Datta has completed a manuscript on hierarchical Bayesian approach based on probit regression to estimate small area proportions. This manuscript is under review for consideration of publication in the *Journal of Survey Statistics and Methodology*. Additionally, Datta is also collaborating with a graduate student and two faculty on two more manuscripts on the Bayesian approach to ranking multiple populations or entities.

Staff: Gauri Datta (x33426), Jerry Maples, Abhyuday Mandal (University of Georgia), Yiren Hou (University of Michigan), Jiacheng Li (Wells Fargo), Aditya Mishra (University of Georgia), David Okech (University of Georgia), Arghadeep Basu (University of Georgia), Hui Yi (University of Georgia)

Spatial Analysis & Modeling

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

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

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

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

Highlights: During FY 2024, there was no significant progress on this project.

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

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

Highlights: During FY 2024, staff developed a nonparametric unit level model for sample survey data collected under an informative sample survey design. This model was extended to account for longitudinal data structures allowing for borrowing of strength over time as well as space. Distributional theory was developed and the methods were coded in an R package. A data set was created using Survey of Income and Program Participation data and the model was used to estimate total wealth at the state level.

Staff: Ryan Janicki (x35725), Paul Parker, Scott Holan (R&M), Daniel Vedensky (University of Missouri & Census Bureau Dissertation Fellow)

C. Machine Learning and Spatial Modeling
Description: The use of auxiliary information such as covariate data and spatial structures in Bayesian hierarchical models is critically important for producing accurate predictions. However, it can often be the case that the quantity of available data is overwhelming, and the number of potential predictors is far greater than the number of observations. In this setting it is challenging to select a manageable subset of predictors for use in a model, to specify a functional form for the relationship between the response and predictor variables, and to include all important interactions and correlations.

Highlights: During FY 2024, staff researched methods for selecting a set of relevant predictor variables from a large collection of auxiliary data sources, but none of the data sources precisely matches the response variable, due to differences in definitions or due to lack of agreement in age, race, or sex categories. Staff utilized an all-to-all method which maps all predictor variables to each response variable. This methodology was applied to American Community Survey and decennial census data to create design matrices for S-DHC county-level response variables. These matrices were used in various regression setups to determine the effectiveness of the proposed methodology.

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

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

Highlights: During FY 2024, staff conducted several exploratory data analysis tasks for applying machine learning methods to a downscaling problem where the target of statistical inference is prediction of the total of a response variable of interest over a user-specified spatial region using a large number of potentially useful covariates. In addition, staff also considered the problem of uncertainty quantification of the prediction, in the form of valid prediction intervals in this context. The ideas are motivated by the GRIDS project, where the observed data are assumed to be aggregated, for example, at the block group level and prediction is sought over a set of regions specified by the (much smaller and non-nested) grid cells. It is observed that a naïve application of the LASSO algorithm does a very poor job of fitting a reasonable regression model that can be directly used at both scales. A novel modification is proposed and is shown, using real data, to produce scale-sensitive predictions with minimal computations. Efforts are underway to (i) validate the LASSO-based prediction methodology theoretically, (ii) develop (model-based) uncertainty quantification methods, and (iii) determine the tuning parameters and finite sample accuracy of the proposed methodology using a detailed simulation study.

Staff: Soumen Lahiri, Ryan Janicki (x35725)

Time Series & Seasonal Adjustment

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

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

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

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

Highlights: During FY 2024, staff explored the enhanced benefits of multivariate signal extraction over the traditional univariate approach, focusing on the potential gains in mean squared error (MSE) reduction when adding additional series to a model. This investigation aimed to assess how much improvement in seasonal adjustment can be achieved using multivariate techniques and included research into Granger-type causality for signal extraction. Additionally, we developed a weekly seasonal modeling framework using a novel differencing operator with a specific unit root structure, tailored for weekly time series with fractional periodicities. This framework was applied to further improve seasonal adjustment methods for such data.

Staff: James Livsey (x33517), Tucker McElroy (R&M), Osbert Pang, Anindya Roy

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

Highlights: During FY 2024, staff developed a methodology that smoothly transitions from univariate to multivariate modeling, maintaining the familiar characteristics of existing models. This approach leverages the Final Equation Form of a VAR specification to integrate multivariate techniques while ensuring a user-friendly experience. Additionally, staff finalized a paper on time series privacy titled *RAFT: A Model Agnostic Utility-Preserving Privacy Mechanism for Time Series Data*, which introduces a filter with random coefficients to privatize time series data while preserving utility.

Staff: James Livsey (x33517), Tucker McElroy (R&M), Osbert Pang, Anindya Roy

Experimentation, Prediction, & Modeling

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

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

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

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

Highlights: During FY 2024, staff submitted a research manuscript studying the impact of the choice of prior distribution on the CMP-parametrized version of the COM-Poisson distribution through theoretical results and data simulations with varying sample sizes. The manuscript has been reviewed for the first round, and staff are revising the manuscript to address reviewer feedback. Staff developed an R package associated with the generalized Conway-Maxwell Poisson (GCMP)

distribution. The package will help analysts to compute various measures of interested associated with the GCMP distribution, including its probability function, cumulative distribution function, quantile function, random number generation, mean, variance, and kth moment. Staff continue to develop documentation associated with the package.

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

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

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

Highlights: During FY 2024, staff has been investigating variance estimates under systematic sampling, including unequal probability systematic sampling. The problem appears to be of considerable interest in view of the extensive use of systematic sampling at the Census Bureau. Variance estimation under systematic sampling has been a challenging problem because an unbiased variance estimate does not exist. Successive difference replication (SDR) has been an approach that has been widely recommended in the literature. Other available variance estimates include those computed after splitting the sample into smaller "systematic samples," and estimates based on an adaptation of the bootstrap to the systematic sampling scenario. Staff compared the expected value of the SDR estimate and that of the estimate computed based on splitting the sample into smaller "systematic samples," with the actual variance of the estimated population mean based on an equal probability systematic sample. This permits the identification of scenarios where either estimate is appropriate, in terms of providing a smaller bias.

Staff plans to conduct numerical studies to compare the different variance estimates and eventually extend the results to unequal probability systematic samples. The bootstrap variance estimate appears to have been less explored, and staff plans to pursue this. Of particular interest will be the case of binary data, where the problem of interest is estimating a population proportion. A primary motivation behind this investigation is to develop appropriate variance estimates that can be recommended in the analysis of embedded experiments that the Demographic Statistical Methods Division is conducting, particularly in evaluating strategies that can improve survey response rates.

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

C. Ratio Edits for Multivariate Data Based on Tolerance Rectangles

Description: Ratio edit tolerances are bounds used for identifying errors in the data obtained by Economic Census Programs so that they can be flagged for further review. The tolerances represent upper and lower bounds on the ratio of two highly correlated items and are used for outlier detection; that is, to identify units that are inconsistent with the rest of the data. When data are bivariate or multivariate, a Mahalanobis distance based outlier detection method has been recommended in the literature. However, this may not adequately flag the outliers, since the outlyingness of a single variable (or a few variables) may be cancelled out by the magnitudes of the other variables. It appears that a rectangular tolerance region that provides simultaneous tolerance intervals on each variable is more appropriate.

Highlights: During FY 2024, staff has been investigating the computation of simultaneous tolerance intervals in order to identify errors in the individual components of a data vector. Even under the parametric setup of a multivariate normal distribution, the computation of such tolerance intervals is challenging since the ratios clearly don't follow a multivariate normal distribution. Thus, staff is exploring a parametric bootstrap approach for computing the required simultaneous tolerance intervals. Of particular interest is the computation of simultaneous central tolerance intervals; that is, intervals that include the central part (say, the central 95% part) of each marginal ratio. Both parametric and non-parametric bootstrap approaches are being pursued for computing the required tolerance limits. The accuracy of the proposed methodologies will be assessed using estimated coverage probabilities. The availability of simultaneous tolerance intervals will be advantageous in identifying erroneous components of an observation vector, instead of declaring the entire data vector to be in error, which is a drawback associated with an ellipsoidal region. The work extends the corresponding univariate results in Young and Mathew (*Journal of Official Statistics, 2015*).

Staff: Thomas Mathew (x35337)

D. Rejection Sampling for Weighted Densities

Description: This project investigates rejection sampling for weighted densities using proposals which relax the weight function. Weighted target distributions arise in

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

Highlights: During FY 2024, staff completed a manuscript on vertical weighted strips (VWS) methodology focusing on univariate target distributions and submitted it to a journal for peer review. Staff received reviewer feedback and began a series of major revisions. One revision includes an application in small area estimation; here, VWS is used to reliably generate from an unfamiliar conditional distribution within a Gibbs sampler.

Staff began development of a 'vws' R package to facilitate implementation of the methodology. The package is intended to provide both an R interface and a C++ interface, where the latter gives improved performance, and safer / more formal type handling and object-orientation.

Staff developed the 'fntl' R package to facilitate work in C++ through Rcpp, including use in the 'vws' package. 'fntl' provides an R-like interface to standard numerical tools in C++ - such as integration, root-finding, and optimization - which take functions as arguments. To make use of the package, functions are coded as C++ lambdas which can be defined on the fly within a program, access variables in the surrounding environment, and be passed as arguments to other functions. An extensive vignette explains the package and provides details about its API. 'fntl' was submitted to CRAN for use by the general R community.

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

Simulation, Data Science, & Visualization

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

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

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

A. Visualizing the United States

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

Highlights: During FY 2024, staff explored visual forms that communicate uncertainty and the full range of datasets to show the population in greater detail. Staff developed and designed visualizations that use interaction and animation to make it easier for individuals to relate to a dataset. This work uses a visual model that starts with individuals and progresses towards aggregates; the approach appears to be beneficial in helping readers understand concepts. Variations on the model will be explored further.

Staff: Nathan Yau (CSRM, FLOWINGDATA.COM)

B. Development and Evaluation of Methodology for Statistical Disclosure Control

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

Highlights: During FY 2024, staff worked on developing suitable randomized response procedures for protecting respondent's privacy and data confidentiality and methods for making valid inferences from randomized response data.

Staff: Tapan Nayak (x35191)

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

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

Highlights: During FY 2024, the following paper published in the open access journal *Mathematics* (ISSN 2227-7390): "Inference about a Common Mean Vector from Several Independent Multinormal Populations with Unequal and Unknown Dispersion Matrices." This paper addresses the problem of making inferences about a common mean vector from several independent multivariate normal populations with unknown and unequal dispersion matrices. Staff and external researchers propose an unbiased estimator of the common mean vector, along with its asymptotic estimated variance, which can be used to test hypotheses and construct confidence ellipsoids, both of which are valid for large samples. Additionally, they discuss an approximate method based on generalized *p*-values. The paper also presents exact test procedures and methods for constructing exact confidence sets for the common mean vector, with a comparison of the local power of these exact tests. The performance of the proposed methods is demonstrated through a simulation study and an application to data from the Current Population Survey (CPS) Annual Social and Economic (ASEC) Supplement 2021 conducted by the U.S. Census Bureau for the Bureau of Labor Statistics.

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

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

Description: This is a continuation of research reported in Klein and Sinha (2015). There is a huge literature on data analysis under privacy or confidentiality protection. Among many inferential statistical methods based on parametric models, data analysis based on perturbation of original sensitive data using plug-in and posterior predictive sampling are quite common. In this paper we consider a very basic inferential problem of tests and confidence intervals for a normal mean with unknown variance based on synthetic data obtained from multiple imputations under posterior predictive sampling method. Several methods are suggested and compared. A general expression of the local power of a class of tests is also derived which can be used in a design context to determine a combination of sample size and number of imputations to guarantee a desired level of local power. A measure of privacy protection is derived to

demonstrate that privacy would be compromised if too many imputations are released. An application to draw inference about the household earnings, corresponding to U.S. Census Bureau data, is illustrated.

Highlights: During FY 2024, staff and external researchers published in *Calcutta Statistical Association Bulletin* the following paper: "Comparison of Tests and Confidence Intervals for Univariate Normal Mean Based on Multiply Imputed Synthetic Data obtained by Posterior Predictive Sampling." There is a huge literature on data analysis under privacy or confidentiality protection. Among many inferential statistical methods based on parametric models, data analysis based on perturbation of original sensitive data using plug-in and posterior predictive sampling are quite common. In this paper, the researchers consider a very basic inferential problem of tests and confidence intervals for a normal mean with unknown variance based on synthetic data obtained from multiple imputations under posterior predictive sampling method. Several methods are suggested and compared. A general expression of the local power of a class of tests is also derived which can be used in a design context to determine a combination of sample size and number of imputations to guarantee a desired level of local power. A measure of privacy protection is derived to demonstrate that privacy would be compromised if too many imputations are released. An application to draw inference about the household earnings, corresponding to a U.S. Census Bureau data, is illustrated.

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

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

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

Highlights: During FY 2024, staff and an external researcher published the following paper in *Sankhya, Series B*: "Analysis of One-Way ANOVA Model using Synthetic Data." In this paper, the authors consider the age–old ANOVA problem of testing the equality of means of several univariate normal populations with a common unknown variance, except that the data used for analysis arise from a synthetic version of the original observations. We address two versions of the synthetic data: one obtained under Plug-In sampling (PIS) method and the other under Posterior Predictive Sampling (PPS) method. We study its distributional properties (null and non-null) and provide enough computational details. A comparison of power is also provided. As expected, the power under the PIS method is more than that under the PPS method. A measure of privacy protection is also evaluated and it turns out that the PIS method provides less protection than the PPS method, thus confirming the standard belief that accuracy of inference and privacy protection work in opposite directions. Robustness of the proposed tests under deviations from normality is also studied.

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

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

Description: In this project we address the problem of constructing a confidence ellipsoid of a multivariate normal mean vector based on a random sample from it. The central issue at hand is the sensitivity of the original data and hence the data cannot be directly used/analyzed. We consider a few perturbations of the original data, namely, noise addition and creation of synthetic data based on the plug-in sampling (PIS) method and the posterior predictive sampling (PPS) method. We review some theoretical results under PIS and PPS which are already available based on both frequentist and Bayesian analysis (Klein and Sinha, 2015, 2016; Guin et al., 2023) and derive the necessary results under noise addition. A theoretical comparison of all the methods based on expected volumes of the confidence ellipsoids is provided. A measure of privacy protection (PP) is discussed and its formulas under PIS, PPS and noise addition are derived and the different methods are compared based on PP. Applications include analysis of two multivariate datasets. The first dataset, with $p = 2$, is obtained from the latest Annual Social and Economic Supplement (ASEC) conducted by the US Census Bureau in 2023. The second dataset, with $p = 3$, pertains to renal variables obtained from the book by Harris and Boyd (1995). Using a synthetic version of the original data generated through PIS and PPS methods and also the noise added data, we produce and display the confidence ellipsoids for the unknown mean vector under various scenarios. Finally, the privacy protection measure is evaluated for various methods and different features.

Highlights: During FY 2024, staff and two external

researchers, received acceptance for the following paper from the *Journal of the Society of Statistics, Computer and Applications (SSCA)* with title: "Confidence Ellipsoids of a Multivariate Normal Mean Vector Based on Noise Perturbed and Synthetic Data with Applications." In this paper, the researchers address the problem of constructing a confidence ellipsoid of a multivariate normal mean vector based on a random sample from it. The central issue at hand is the sensitivity of the original data and hence the data cannot be directly used/analyzed. They consider a few perturbations of the original data, namely, noise addition and creation of synthetic data based on the plug-in sampling (PIS) method and the posterior predictive sampling (PPS) method. They review some theoretical results under PIS and PPS which are already available based on both frequentist and Bayesian analysis (Klein and Sinha, 2015, 2016; Guin et al., 2023) and derive the necessary results under noise addition. A theoretical comparison of all the methods based on expected volumes of the confidence ellipsoids is provided. A measure of privacy protection (PP) is discussed and its formulas under PIS, PPS and noise addition are derived and the different methods are compared based on PP. Applications include analysis of two multivariate datasets. The first dataset, with $p = 2$, is obtained from the latest Annual Social and Economic Supplement (ASEC) conducted by the U.S. Census Bureau in 2023. The second dataset, with $p = 3$, pertains to renal variables obtained from the book by Harris and Boyd (1995). Using a synthetic version of the original data generated through PIS and PPS methods and also the noise added data, we produce and display the confidence ellipsoids for the unknown mean vector under various scenarios. Finally, the privacy protection measure is evaluated for various methods and different features.

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

Summer at Census

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

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

Highlights: During FY 2024, staff organized the fifteenth annual *2024 SUMMER AT CENSUS* which brought 12 recognized scholars to the Census Bureau for 1-3 day (virtual) visits and cover broad themes including: data quality and statistical inference, disclosure avoidance, large language models, marketing and public outreach, measurement of race and ethnicity, probability sampling, record linkage, synthetic social habitats, time series, and usability/human-computer interactions. Each scholar engaged in collaborative research with Census Bureau researchers and staff (Center for Statistical Research & Methodology, Center for Optimization & Data Science, Center for Enterprise Dissemination, Research & Methodology Directorate, Economic Statistical Methods Division, and Center for Behavioral Science Methods) on at least one current specific Census Bureau problem and presented a seminar based on his/her research.

Staff: Tommy Wright (x31702), Joseph Engmark

Research Support and Assistance

This staff provides substantive support in the conduct of research, research assistance, technical assistance, and secretarial support for the various research efforts.

Staff: Joseph Engmark, Michael Hawkins, Kelly Taylor

3. PUBLICATIONS

3.1 JOURNAL ARTICLES (Peer-Reviewed), PUBLICATIONS

Aleshin-Guendel, S. and Steorts, R. (2024). "Monitoring Convergence Diagnostics for Entity Resolution," *Annual Review of Statistics and Its Applications, Vol 11, 419-435*. https://doi.org/10.1146/annurev-statistics-040522-114848.

Aleshin-Guendel, S. and Wakefield, J. (2024). "Adaptive Gaussian Markov Random Fields for Child Mortality Estimation," *Biostatistics*, kxae030. https://doi.org/10.1093/biostatistics/kxae030.

Aleshin-Guendel, S., Sadinle, M., and Wakefield, J. (2024). "The Central Role of the Identifying Assumption in Population Size Estimation," *Biometrics* (with Discussion), *80 (1)*, viad028.

Basak, B. and Sinha, B. (2024). "Analysis of One-way ANOVA Model Using Synthetic Data," *Sankhya (B), Volume 86*, 164-190.

Basak, B., Yehenew, G.K., and Sinha, B.K. (In Press). "Confidence Ellipsoids of a Multivariate Normal Mean Vector Based on Noise Perturbed and Synthetic Data with Applications," *Journal of Society of Statistics, Computer and Applications (SSCA), Special Issue Dedicated to the Fond Memories of Prof C.R. Rao on "Life and Work of C.R. Rao (1920-2023): The Revolutionary of Statistical Sciences*, *Vol. 22*.

Basak, B. and Sinha, B. (In Press). "Comparison of Tests and Confidence Intervals for Univariate Normal Mean Based on Multiply Imputed Synthetic Data Obtained by Posterior Predictive Sampling," *Calcutta Statistical Association Bulletin*.

Ben-David, E., West, B.T., and Slawski, M. (2023). "A Novel Methodology for Improving Applications of Modern Predictive Modeling Techniques to Linked Data Sets Subject to Mismatch Error," *2023 Big Data Meets Survey Science (BigSurv)*, Quito, Ecuador, 2023, pp. 1-8, doi: 10.1109/BigSurv59479.2023.10486610

Bukke, P., Ben-David, E., Diao, G, Slawski, M., and West, B.T. (In Press). "Cox Proportional Hazards Regression Using Linked Data: An Approach Based on Mixture Modelling," *IISA Series on Statistics and Data Science*, to appear. Winner of the 2024 ASA GSS/SRMS/SSS Student Paper Award.

Datta, G.S., Lee, J., and Li, J. (2024). "Pseudo-Bayesian Small Area Estimation," *Journal of Survey Statistics and Methodology, Vol. 24*, 343-368. https://doi.org/10.1093/jssam/smad012

Datta, G.S. and Li, J. (In Press). "A Quasi-Bayesian Approach to Small Area Estimation Using Spatial Models," *Calcutta Statistical Association Bulletin.* https://doi.org/10.1177/00080683231199021

Deo, N., Sanguthevar R., Joyanta B., Soliman, A., Weinberg, D., and Steorts, R. (2023). "Novel Blocking Techniques and Distance Metrics for Record Linkage," *Proceedings of The 25th International Conference on Information Integration and Web Intelligence (iiWAS), Lecture Notes in Computer Sciences*, Springer, 431-446.

Guin, A., Roy, A., and Sinha, B. (2023). "Bayesian Analysis of Singly Imputed Synthetic Data under the Multivariate Normal Model," *International Journal of Statistical Sciences, Volume 23(2)*, 1-18.

Ibrahim, S., Mazumder, R., Radchenko, P., and Ben-David, E. (In Press). "Predicting Census Survey Response Rates with Parsimonious Additive Models and Structured Interactions," *The Annals of Applied Statistics.*

Janicki, R., Holan, S. H., Irimata, K. M., Livsey, J., and Raim, A. (2023). "Spatial Change of Support Models for Differentially Private Decennial Census Counts of Persons by Detailed Race and Ethnicity," *Journal of Statistical Theory and Practice, Vol 17, 13 (2023).* https://doi.org/10.1007/s42519-023-00328-5.

Joyce, P.M. and McElroy, T.S. (In Press). "Modeling Survey Time Series Data with Flow-Observed CARMA Processes," *Journal of Official Statistics*.

Kang, J., Morris, D.S., Joyce, P., and Dompreh, I. (2023). "On Calibrated Inverse Probability Weighting and Generalized Boosting Propensity Score Models for Mean Estimation with Incomplete Survey Data," *Wiley Interdisciplinary Reviews (WIREs) Computational Statistics, Vol 15, No. 6*.

Kaputa, S. J., Morris, D.S., and Holan S.H. (2024). "Bayesian Multisource Hierarchical Models with Applications to the Monthly Retail Trade Survey," *Journal of Survey Statistics and Methodology,* https://doi.org/10.1093/jssam/smae019.

Kifle, Y.G., Moluh, A.M., and Sinha, B.K. (In Press). "Inference about a Common Mean Vector from Several Independent Multinormal Populations with Unequal and Unknown Dispersion Matrices," *Mathematics, 12(17).*

Livsey, J. and McElroy, T. (In Press). "Applying the Expectation-Maximization Algorithm to Multivariate Signal Extraction," *Journal of Official Statistics*.

McElroy, T., Ghosh, D., and Lahiri, S. (2023). "Quadratic Prediction of Time Series via Autocumulants," *Sankhya A*, *Vol 86(1)*, 431-463.

McElroy, T. and Politis, D. (2024). "Estimating the Spectral Density at Frequencies Near Zero," *Journal of the American Statistical Association*, *vol. 119 (545)*, 612-624.

Mulry, M.H., Tello-Trillo, C.J., Mule, V.T., and Keller, A. (2024). "Comparisons of Administrative Records Rosters to Census Self Responses and Nonresponse Follow-up Responses," *Statistical Journal of the IAOS, Vol. 40, No. 1,* 41-52.

Parker, P., Holan, S.H., and Janicki, R. (2024). "Conjugate Modeling Approaches for Small Area Estimation with Heteroscedastic Structure," *Journal of Survey Statistics and Methodology, Vol 12,* 1061 – 1080.

Slawski, M., West, B.T., Bukke, P., Wang, Z., Diao, G., and Ben-David, E. (In Press). "A General Framework for Regression with Mismatched Data Based on Mixture Modeling," *Journal of the Royal Statistical Society Series A.*

Slud, E., Hall, A., and Franco, C. (2024). "Small Area Estimates for Voting Rights Act Section 203(b) Coverage Determinations," *Calcutta Statistical Association Bulletin, 76(1),* 137-159, published online 2/13/24, available at https://doi.org/10.1177/00080683231215985.

Su, X., Quaye, G., Wei, Y., Kang, J., Liu, L., Yang, Q., Fan, J., and Levine, R. (In Press). "Smooth Sigmoid Surrogate (SSS): An Alternative to Greedy Search in Decision Trees," *Mathematics, 12,* 3190, https://doi.org/10.3390/math12203190.

West, B.T., Slawski, M., and Ben-David, E. (In Press). "Improved Ensemble Predictive Modeling Techniques for Linked Social Media and Survey Data Sets Subject to Mismatch Error," *Methods, Data, Analyses (MDA).*

Wright, T. (In Press). "Optimal Tightening of the KWW Joint Confidence Region for a Ranking," *Statistics and Probability Letters.*

3.2 BOOKS/BOOK CHAPTERS

3.3 PROCEEDINGS PAPERS

3.4 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY RESEARCH REPORT SERIES
https://www.census.gov/topics/research/stat-research/publications/working-papers/rrs.html

RR (Statistics #2024-01): Tommy Wright, "Understanding and Optimal Tightening of the KWW Joint Confidence Region for a Ranking," February 12, 2024.

RR (Statistics #2024-02): Tucker S. McElroy, Osbert C. Pang, and Baoline Chen, "Mitigating Residual Seasonality while Preserving Accounting Relations in Hierarchical Time Series," February 12, 2024.

RR (Statistics #2024-03): Tommy Wright, "A Joint Confidence Region for a Ranking Based on Differences," June 4, 2024.

RR (Statistics #2024-04): Biswajit Basak, Yehenew G. Kifle, and Bimal K. Sinha, "Confidence Ellipsoids of a Multivariate Normal Mean Vector Based on Noise Perturbed and Synthetic Data with Applications," June 27, 2024.

RR (Statistics #2024-05): Thibaudeau, Yves, "A Review of Modern Multinomial-Derived and Partition-Based Record-

Linkage Methods," August 15, 2024.

RR (Statistics #2024-06): Biswajit Basak and Bimal Sinha, "Comparison of Tests and Confidence Intervals for Univariate Normal Mean Based on Multiply Imputed Synthetic Data Obtained by Posterior Predictive Sampling," September 5, 2024.

RR (Computing #2024-01): Andrew Raim, "fntl: Numerical Tools for Rcpp and Lambda Functions," July 17, 2024.

3.5 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY STUDY SERIES

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

SS (Statistics #2024-01): Andrew Raim, Renee Ellis, and Mikelyn Meyers, "A Multinomial Analysis of Bilingual Training and Nonresponse Follow-up Contact Rates in the 2020 Decennial Census," June 3, 2024.

3.6 OTHER REPORTS

4. TALKS AND PRESENTATIONS

Department of Art and Sciences Seminar, Washington University, St. Louis, November 14, 2023.

- Emanuel Ben-David (Invited Talk), "Improving Applications of Modern Predictive Modeling Techniques to Linked Data Sets Subject to Mismatch Error."

Department of Mathematics Seminar, Pohang Institute of Science and Technology (POSTECH), South Korea, December 7, 2023.

- Joseph Kang, "On Machine Learning Models for Incomplete Survey Data."

Department of Statistics Seminar, Korea University, South Korea, December 14, 2023.

- Joseph Kang, "Causal Inference of Latent Classes in Complex Survey Data with the Estimating Equation Framework."

Department of Statistics Seminar, Korea University, South Korea, December 15, 2023.

- Joseph Kang, "Estimating the Causal Effects of Risky Behaviors on Herpes."

Department of Statistics Seminar, Seoul National University, South Korea, December 15, 2023.

- Joseph Kang, "Causal Inference of Latent Classes in Complex Survey Data with the Estimating Equation Framework."

Statistics Department Seminar, University of California, Santa Cruz, January 29, 2024.

- Ryan Janicki (Invited Talk, virtual), "Spatial Change of Support Models for Differentially Private Decennial Census Counts of Persons by Detailed Race and Ethnicity."

Department of Statistical Sciences Colloquium, Wake Forest University, March 26, 2024.

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

The Annual General Meeting for 2023-2024 of the Calcutta Statistical Association, March 28, 2024.

- Gauri S. Datta (Invited Talk, virtual), "Credible Distributions for Ranking of Entities."

Workshop on "Discrete Distributions" in Memory of A. F. Kemp, Athens, Greece, April 13, 2024.

- Kimberly Sellers (Invited Speaker, virtual), "Bivariate COM-Poisson Distributions."

15th Annual Probability & Statistics Day, University of Maryland, Baltimore County, April 20, 2024.

- Tommy Wright (Invited Banquet Speaker), "Visualizations & Uncertainty."

The IAOS-ISI 2024, Improving Decision Making for All, Mexico City, Mexico, May 15-17, 2024.

- Gauri S. Datta (Invited Talk), "Credible Distributions for Ranking of Entities."

Small Area Estimation 2024 Conference, *Celebrating the 65th Birthday of Prof. Partha Lahiri*, Pontificia Universidad Católica del Perú, Lima, Peru, June 3-7, 2024.

- Emanuel Ben-David (Invited Talk), "Improving Applications of Modern Predictive Modeling Techniques to Linked Data Sets Subject to Mismatch Error". (Invited Short Course): "Data Analysis with Linked Datasets."
- Gauri S. Datta (Invited Talk), "Credible Distributions for Ranking of Entities."
- Jerry Maples (Invited Talk), "Using Geographically Weighted Regressions to Assess Variability in Small Area Model Parameters."
- Eric Slud (Invited Talk), "Evaluating Weight Adjustments Using Short Time Series of Survey Estimates at Multiple Geographic Levels."

2024 NSF Summer Program in Research and Learning/Summer Program Advancing Techniques in the Applied Learning of Statistics Colloquium, Georgetown University, Washington, D.C., June 14, 2024.

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

International Conference on Establishment Statistics, Glasgow, Scotland, June 17-20, 2024.

- Darcy Morris, "Ratio-Synthetic and Imputation-Based Models Using Integrated Survey and Alternative Data Sources for Retail Sales Estimates."

- James Livsey (Invited Short Course, Chair), "Seasonal Adjustment & Time Series Analysis."

Conference for African American Researchers in the Mathematical Sciences 2024, Tufts University, Medford, MA, June 28, 2024.

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

Recent Trends in Statistical Theory and Applications (WSTA 2024, virtual), University of Kerala, India, June 29, 2024.

- Gauri S. Datta (Plenary Talk), "Small Area Estimation with Non-normal Random Effects."

2024 International Society for Bayesian Analysis (ISBA) World Meeting, Venice, Italy, July 1-7, 2024.

- Adam Hall, "Bayesian Multinomial Logit Model for the Voting Rights Act, Section 203."

2024 NSF Summer Program in Research and Learning/Summer Program Advancing Techniques in the Applied Learning of Statistics Colloquium, Georgetown University, Washington, D.C., July 12, 2024.

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

2024 Joint Statistical Meetings, American Statistical Association, Portland, Oregon, August 3-8, 2024.

- Serge Aleshin-Guendel, "The Central Role of the Identifying Assumption in Population Size Estimation."
- Emanuel Ben-David, "Improving Applications of Modern Prediction Modeling Techniques to Linked Data Sets Subject to Mismatched Error."
- Gauri Datta, "Credible Distributions of Overall Ranking of Estimates."
- Adam Hall, "Bounding the Error of the Maximum Ratio Test of Unacceptability."
- Kyle Irimata, "Small Area Modeling for Differentially Private Counts."
- Patrick Joyce, "Statistical Methods Underlying International Migration Estimates at the Census Bureau."
- James Livsey, "Error Reduction in Multivariate Signal Extraction."
- Darcy Morris, "Multi-Source Hierarchical Models for Geographically Granular Retail Sales Estimates."
- Mary Mulry, "Some Results from the Continuous Count Study."
- Paul Parker, "Statistical Deep Learning for Dependent Establishment Data."
- Andrew Raim, "Rejection Sampling with Vertical Weighted Strips."
- Anindya Roy, "Multi-step Ahead Forecasting Using Transformer Based Model."
- Eric Slud, "Assessment of Effectiveness of Weighting Adjustment Using Short Series of Survey Estimates of Multiple Geographic Levels."
- Yves Thibaudeau, "Recent Bayesian and Non-Bayesian Evaluating Methods for Analyzing Sparse Classifications."
- Yves Thibaudeau, "Computer Science and Statistical Theory in Record Linkage: Will One Continue to Help the Other?"
- Daniel Weinberg, "Triple System Estimation of National Population Counts Through Log-linear and Latent Class Modeling."
- Tommy Wright, "A New Joint Confidence Region for a Ranking."

2024 Royal Statistical Society (RSS) International Conference, Brighton, UK, September 2-5, 2024.

- Joseph Kang, "Bayesian Multinomial Logit Models for the U.S. Voting Rights Act (VRA) Determination."

The 2024 International Conference on Trends and Perspectives in Linear Statistical Inference, LinStat'2024, Poprad, Slovakia, September 2-6, 2024.

- Emanuel-Ben-David (Invited Talk) "Gaussian DAG Models with Imposed Symmetries."

Department of Mathematics and Statistics Seminar, Indian Institute of Science Education and Research (IISER) Kolkata, West Bengal, India, September 6, 2024.

- Bimal Sinha, "Confidence Ellipsoids of a Multivariate Normal Mean Vector Based on Noise Perturbed and Synthetic Data with Applications."

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

Dan Weinberg (Joint work with Yves Thibaudeau), U.S. Bureau of the Census, "Using Orthogonalized Design Matrices to Improve Estimation of Record Linkage Parameters," February 13, 2024.

Gauri Datta, University of Georgia/U.S. Census Bureau, "Credible Distributions for Ranking Entities," February 27, 2024.

Nathan Yau, FLOWINGDATA.COM/U.S. Census Bureau, "Balancing Between Noise and Signal in Visualization," March 26, 2024.

Bimal Sinha, University of Maryland, Baltimore County/U.S. Census Bureau, "Confidence Ellipsoids of a Multivariate Normal Mean Vector Based on Noise Perturbed and Synthetic Data with Applications," April 2, 2024.

Emanuel Ben-David, U.S. Census Bureau, "Improving Applications of Predictive Modeling with Linked Data Sets," May 7, 2024.

Madhav Marathe & Samarth Swarup, University of Virginia, *SUMMER (Virtually) AT CENSUS*, "Synthetic Social Habitats for Policy and Decision Making," May 29, 2024.

Martin Slawski, George Mason University, *SUMMER (Virtually) AT CENSUS*, "Some Recent Advances and Open Problems in Post-Linkage Data Analysis," June 11, 2024.

Yao Zheng, University of Connecticut, *SUMMER (Virtually) AT CENSUS*, "New Advances in High-Dimensional Time Series Modeling," June 12, 2024.

Priyanjali Bukke & Martin Slawski, George Mason University, *SUMMER (Virtually) AT CENSUS*, "An Introduction to R Package "pldamixture" for Post-Linkage Data Analysis," June 13, 2024.

Kimberly Huyser, The University of British Columbia & Desi Small-Rodriguez, University of California, Los Angeles, *SUMMER (Virtually) AT CENSUS*, "Honoring Indigenous Lives and Life Experiences in Data," June 18, 2024.

Bikas Sinha, Retired Professor of Statistics, Indian Statistical Institute, Kolkata, *SUMMER (Virtually) AT CENSUS*, "Understanding Features of Probability Proportioned to Size Sampling Designs (both with and without replacement) and Properties of Unbiased Estimators of a Population Mean," June 20, 2024.

Trent Buskirk, Old Dominion University, *SUMMER (Virtually) AT CENSUS*, "Can We Count on Chatbots at the Census Bureau? Exploring the Possibility and Implausibility of Large Language Models in Survey Science," June 25, 2024.

Soussan Djamasbi, Worchester Polytechnic Institute, *SUMMER (Virtually) AT CENSUS*, "Driving Innovation through User Experience Research," July 10, 2024.

Brian Liu, Massachusetts Institute of Technology (MIT), "Making Tree Ensembles Interpretable," July 16, 2024.

Ashwin Machanavajjhala, Tumult Labs, *SUMMER (Virtually) AT CENSUS*, "Privately Tabulating Skewed Magnitude Variables with Per Record Differential Privacy," July 16, 2024.

Shunyuan Zhang, Harvard University, *SUMMER (Virtually) AT CENSUS*, "Ambiguity in Multi-modal Digital Ads," July 17, 2024.

Dennis K.J. Lin, Purdue University, *SUMMER (Virtually) AT CENSUS*, "Artificial Intelligence, Biological Intelligence, Statistical Intelligence," July 18, 2024.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

Fulbright-Nehru Student Research Fellow

- **Joseph Engmark –** to study at the Indian Statistical Institute in Kolkata, West Bengal (INDIA). The fellowship began in late September 2024 and ends in June 2025. He will be hosted and advised by Diganta Mukkerjee, Sampling and Official Statistics Unit. The proposed project is research on capture-recapture methodology. This research will contribute to Joe's PhD dissertation in Mathematical Statistics at the University of Maryland, College Park.

6.2 SIGNIFICANT SERVICE TO PROFESSION

Serge Aleshin-Guendel

- Reviewer, Grant Proposal, National Science Foundation
- Refereed papers for *Annals of Applied Statistics, Journal of the Royal Statistical Society Series A, and Statistical Methods & Application.*

Emanuel Ben-David

- Refereed papers for the *IEEE Transactions on Information Theory*
- Member, 2024 W.J. Dixon Award for Excellence in Statistical Consulting Committee, American Statistical Association
- Member, Program Committee, 17th (2024) International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation

Gauri Datta

- Organizer, Invited Session, 2024 SAE Conference, Lima, Peru
  Instructor, Short Course (2 hr): Bayesian Methods in Small Area Estimation, 2024 SAE Conference, Lima, Peru
  -This course exposed existing Bayesian methods in SAE to practitioners of survey sampling from many Latin American countries who attended the conference.
- Associate Editor, *Sankhya*
- Associate Editor, *Journal of the Royal Statistical Society, Series A*
- Associate Editor, *Environmental and Ecological Statistics*
- Editorial Member, *Calcutta Statistical Association Bulletin*
- Refereed papers for *Journal of the Royal Statistical Society A, Sankhya, Survey Methodology, Journal of Survey Statistics and Methodology, Statistics and Probability Letters, Journal of Applied Statistics, Journal of the American Statistical Association*

Kyle Irimata

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

Ryan Janicki

- Refereed a paper for *Journal of the Royal Statistical Society, Series A*
- Refereed a paper for *Journal of Survey Statistics and Methodology*
- Member, PhD Qualifying Exam Committee, Statistics Department, University of California, Santa Cruz

Patrick Joyce

- Refereed a paper for *Regional Statistics*

Joseph Kang

- Associate Editor, *Journal of Addiction and Prevention*
- Refereed papers for *Statistical Methods in Medical Research*

James Livsey

- Member, PhD in Statistics Committee, University of California, Santa Cruz
- Member, PhD in Statistics Committee, Mississippi State University

Jerry Maples

- Refereed paper for *Journal of the Royal Statistical Society – Series A, Journal of Official Statistics*
- Member, Expert Panel for X-59 Community Response Testing, NASA Langley Research Center (Hampton, VA)

Thomas Mathew

- Associate Editor, *Sankhya*
- Associate Editor, *Journal of Multivariate Analysis*
- Associate Editor, *Journal of Occupational and Environmental Hygiene*
- Refereed papers for *Journal of Statistical Computation and Simulation, Journal of the Royal Statistical Society, Series C, Pharmaceutical Statistics, Statistical Methods in Medical Research, Statistics in Medicine*, and *BMC Medical Research Methodology*
- Member, W.J. Youden Award in Interlaboratory Testing Committee, American Statistical Association

Darcy Morris

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

Mary Mulry

- Associate Editor, *Journal of Official Statistics*
- Refereed a paper for *Journal of the Royal Statistics Society - A*

Tapan Nayak

- Associate Editor, *Journal of Statistical Theory and Practice*
- Guest Editor, Special Issue of *Statistics and Applications* in Memory of Professor C.R. Rao
- Refereed papers for *Journal of Official Statistics, Journal of Survey Statistics and Methodology, Applied Stochastic Models in Business and Industry*

Kimberly Sellers

- Associate Editor, *The American Statistician*
- Guest Editor, *Applied Stochastic Models in Business and Industry*
- Past Chairperson (1/1/23-12/31/23) Justice, Equity, Diversity, and Inclusion (JEDI) Outreach Group, American Statistical Association
- Chairperson, External Nominations and Awards Committee, American Statistical Association
- Member, Committee on Applied and Theoretical Statistics, National Academies of Sciences, Engineering, & Medicine
- Reviewer, *Scientific Reports*

Bimal Sinha

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

Eric Slud

- Associate Editor, *Journal of Survey Statistics and Methodology*
- Associate Editor, *Lifetime Data Analysis*
- Associate Editor, *Statistical Theory and Related Fields*
- Refereed papers for *Metron, Annals of Statistics*, *Sankhya B*, *Journal of Official Statistics, Journal of American Statistical Association, Survey Methodology*
- Organizer, Session--*"Assessment of Survey Weight-Adjustment Methods,"* Small Area Estimation 2024 Conference, Lima, Peru

Rebecca Steorts

- Associate Editor, *Journal of Survey Statistics and Methodology*

- Associate Editor, *Journal of the American Statistical Association, Applications and Case Studies*
- Associate Editor, *Science Advances*
- Associate Editor*, Bayesian Analysis*

Tommy Wright
- Refereed a paper for *Chance*
- Reviewer, Tenure Review of Faculty Member (2023), Department of Statistics, Colby College

6.3 PERSONNEL NOTES

Shane Lubold accepted a position and left the Census Bureau**.**

Sixia Chen (Statistics Faculty at University of Oklahoma, Health Sciences Center) accepted an ASA/NSF/Census Research Fellowship (two years) to work on "Developing Analytic Tools for Handling Nonprobability Samples with Application to Census Bureau Data Files."

APPENDIX A **Center for Statistical Research and Methodology FY 2024**
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		
6550J08	Data Products Dissemination Preparation/Review/Approval		
5350K04	Demographic Frame 2030		
5450K21	Response Data Quality 2030		
5450K23	Response Processing Planning & Support		
5550K02	Redistricting Data Program		
5650K01	Evaluations, Experiments, & Research		
5650K02	Post Enumeration Survey Design & Estimation		
	1. Summary of Administrative Records Modeling for Some Enumerations in the 2020 Census	Mary Mulry	Tom Mule
	2. Comparisons of Administrative Records Rosters to Census Self-Responses and NRFU Household Member Responses	Mary Mulry	Tom Mule
	3. Supplementing and Supporting Non-Response with Administrative Records	Michael Ikeda	Tom Mule
	4. Experiment for Effectiveness of Bilingual Training	Andrew Raim	Mikelyn Meyers
	5. Coverage Measurement Research	Jerry Maples	Tim Kennel
	6. Empirical Investigation of the Minimum Total Population of a Geographic District to Have Reliable Characteristics of Various Demographic Groups	Kyle Irimata	James Whitehorne
	7. Statistical Modeling to Augment 2020 Disclosure Avoidance System	Andrew Raim/Ryan Janicki	Michael Walsh
	8. Imputation Modeling for Multivariate Categorical Characteristic Data in 2030 Census	Darcy Morris	Tom Mule
	9. 2030 Characteristic Imputation Modeling Research for Nested Data with Structural Zeros	Emanuel Ben-David	Tom Mule
	10. Mobile Questionnaire Assistance: Analysis and Simulation	Andrew Raim	Lisa Moore
	11. Agreements for Advancing Record Linkage	Rebecca Steorts	Krista Park
	12. Continuous Count Study	Michael Ikeda	Tom Mule
	13. Capture-Recapture Coverage Measurement using Administrative Records (Continuous Count Study)	Dan Weinberg	Tom Mule
	14. Cohort Component Birth Modeling (Continuous Count Study)	Emanuel Ben-David	Tom Mule
	15. Cohort Component Domestic Migration Modeling (Continuous Count Study)	Ryan Janicki	Tom Mule
	16. Record Linkage Support for Decennial Census	Ned Porter	Aaron Gilary
6385K70	American Community Survey (ACS)		
	17. Voting Rights ACT (VRA) Section 203 Research Towards 2026 Determinations (also Decennial Project 6550J06)	Joseph Kang	James Whitehorne
	18. Assessing and Enhancing the ACS Experimental Weighting Approach Implemented in 2020 Data Products	Darcy Morris	Mark Asiala
	DEMOGRAPHIC		
TBA	Demographic Statistical Methods Division (DSMD) Special Projects		
	19. Research on Biases in Successive Difference Replication Variance Estimation in Very Small Domains	Eric Slud	Tim Trudell
	20. Nonresponse Adjustments for High Frequency Low Response Surveys	Darcy Morris	John Chestnut
0906/1444X00	Demographic Surveys Division (DSD) Special Projects		
	21. Using Machine Learning for Improving Nonresponse Adjustment in the Current Population Survey	Emanuel Ben-David	Tim Trudell
	22. Data Integration	Ned Porter	Chris Boniface
7165024	Social, Economic, and Housing Statistics Division Small Area Estimation Projects		
	23. Research for Small Area Income and Poverty Estimates (SAIPE)	Jerry Maples	Wes Basel
	24. Assessing Constant Parameters across Areas in the SAIPE Models	Jerry Maples	Wes Basel

	ECONOMIC	
1183X01	General Economic Statistical Support	
1183X90	General Economic Statistical Program Management	
	25. Use of Big Data for Retail Sales Estimates.....	Darcy Morris.....Stephen Kaputa
	26. Seasonal Adjustment Support.....	James Livsey..... Kathleen McDonald-Johnson
	27. Seasonal Adjustment Software Development and Evaluation	James Livsey..... Kathleen McDonald-Johnson
	28. Research on Seasonal Time Series - Modeling & Adjustment Issues.....	James Livsey..... Kathleen McDonald-Johnson
	29. Supporting Documentation & Software for Seasonal Adjustment	James Livsey..... Kathleen McDonald-Johnson
	30. Exploring New Seasonal Adjustment & Signal Extraction Methods.....	James Livsey.....Colt Viehdorfer
	31. Small Area Estimation for the Annual Integrated Economic Survey.....	Serge Aleshin-Guendel .....Stephen Kaputa
	PROGRAM DIVISION OVERHEAD	
0331000	*32. Research Computing*.....	Chad Russell ..... Jaya Damineni

APPENDIX B

FY 2024 PROJECT PERFORMANCE MEASUREMENT QUESTIONNAIRE

CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY

Dear

As a sponsor for the FY 2024 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 2024 ***(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 2024 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 2024 met expectations.

□ Strongly Agree
□ Agree
□ Disagree
□ Strongly Disagree

5. Please provide suggestions for future improved communications or any area needing attention on this project or subproject.

_____
Sponsor Contact Signature Date

Center for Statistical Research and Methodology

Research & Methodology Directorate

STATISTICAL COMPUTING AREA

Joseph Kang (Acting)

Record Linkage & Machine Learning Research Group

Yves Thibaudeau
- Emanuel Ben-David
- Xiaoyun Lu
- Rebecca Steorts
- Dan Weinberg

Missing Data & Observational Data Modeling Research Group

Darcy Morris
- Isaac Dompreh
- Jun Shao (U. of WI)

Joseph Kang

Research Computing Systems & Applications Group

Chad Russell
- Tom Petkunas
- Ned Porter

Simulation, Data Science, & Visualization Research Group

Tommy Wright (Acting)
- Bimal Sinha (UMBC)
- Nathan Yau (FLOWINGDATA.COM)

MATHEMATICAL STATISTICS AREA

Eric Slud

Sampling Estimation & Survey Inference Research Group

Eric Slud (Acting)
- Sixia Chen (ASA/NSF/Census Research Fellow)
- Mike Ikeda
- Patrick Joyce
- Mary Mulry
- Tapan Nayak (GWU)

Small Area Estimation Research Group

Jerry Maples
- Gauri Datta
- Kyle Irimata

Spatial Analysis & Modeling Research Group

Ryan Janicki
- Serge Aleshin-Guendel
- Soumendra Lahiri (Washington U.)
- Paul Parker (U. of CA, Santa Cruz)

Time Series & Seasonal Adjustment Research Group

James Livsey
- Osbert Pang
- Anindya Roy (UMBC)

Experimentation, Prediction, & Modeling Research Group

Tommy Wright (Acting)
- Thomas Mathew (UMBC)
- Andrew Raim
- Kimberly Sellers (NC State U.)

OFFICE OF THE CHIEF

Tommy Wright
- Kelly Taylor
- Joe Engmark
- Adam Hall
- Michael Hawkins