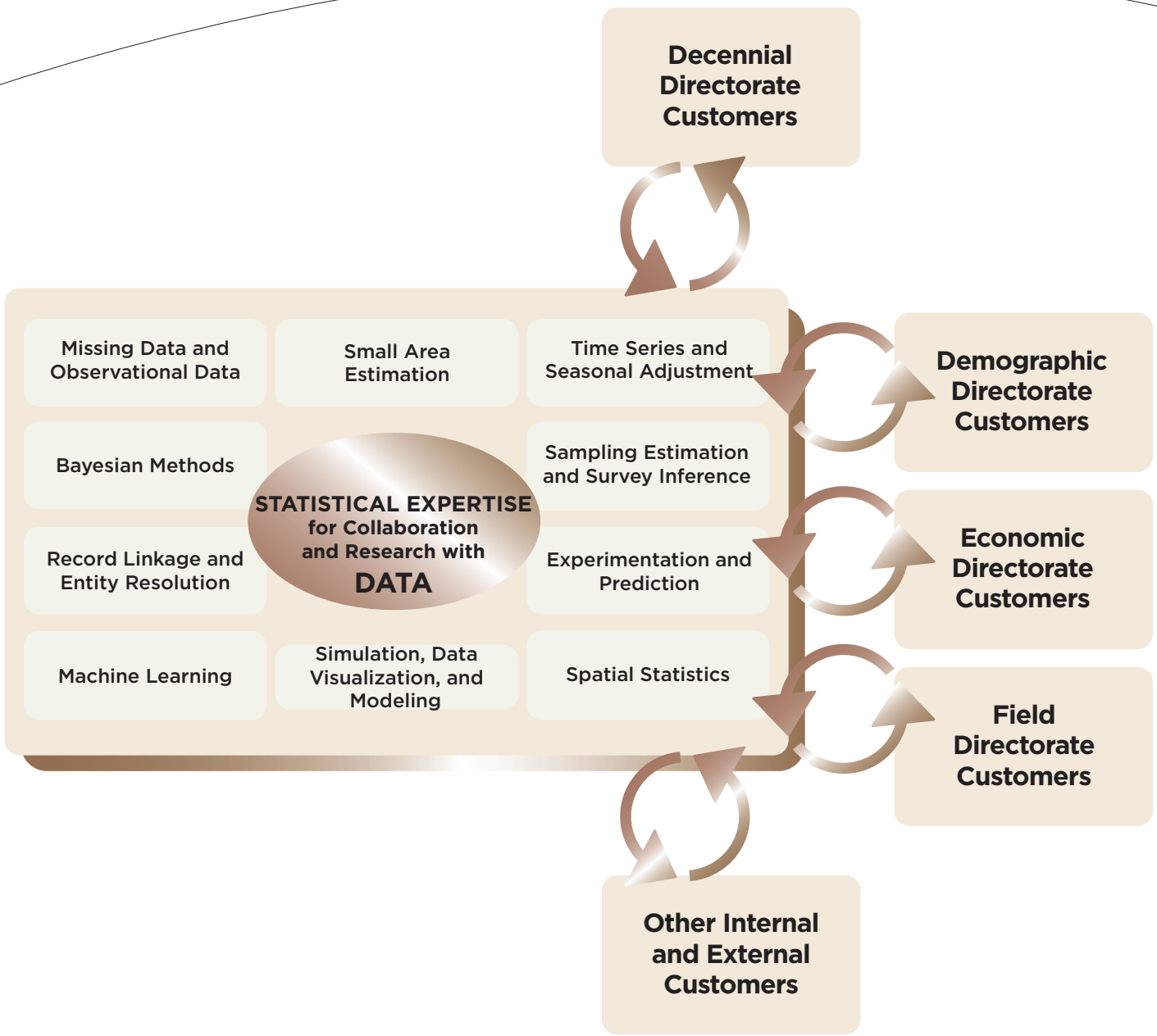
Annual Report of the Center for Statistical Research and Methodology

Research and Methodology Directorate

Fiscal Year 2025



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 2025, 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 (CSRM) staff members made contributions during FY 2025 follow, and more details are provided within subsequent pages of this report:

- CSRM researchers and colleagues in the Research & Methodology Directorate and MITRE produced a research report (under internal final review) detailing the use of Bayesian regression models to produce statistically post-processed population counts for all counties in the United States using data from a Disclosure Avoidance System (DAS). [CSRM (Irimata, Raim, Janicki, Livsey, Hall, Maples); R&M (Holan); MITRE]
- A CSRM researcher and colleagues in the Decennial Census Management Division developed several R packages that implement modeling ideas including: (1) a panel count model that captures tract-level response rates over intervals of time within a census and (2) functions to compute a spatio-temporal "proximity" metric that associates tracts with times and locations of Mobile Questionnaire Assistance (MQA) events that occurred in the vicinity. [CSRM (Raim); DCMD (Moore, Parker)]
- CSRM researchers completed a manuscript detailing the use of regression models (linear and BART) to predict block group level reliability for the 2020 PL94-171 data. This project is an attempt to follow up on some previously observed empirical phenomenon and to take a closer look, using statistical modeling, at variability for smaller districts and to seek an answer to the question "What is the minimum Total (Ideal) population of a district to have reliable characteristics of various demographic groups"? [CSRM (Irimata, Wright)]
- CSRM researchers: (1) completed research comparing results of fitting statistical models to direct estimates of state-level in migration, out migration, and net migration by 4 race/ethnicity categories crossed with 2 age categories; (2) created code to fit multivariate Fay-Herriot model to bivariate direct estimates of in, out, and net migration for each state and again for 4 race/ethnicity categories and 2 age categories; and (3) started a report documenting this work. [CSRM (Janicki, Aleshin-Guendel); R&M (Mule); POP (Jensen, Miller)]
- CSRM researchers and colleagues in the Decennial Statistical Studies Division, the Center for Economic Studies, and the Social, Economic, and Housing Statistics Division: (1) finalized a study of the experimental weighting methodology (in response to unprecedented levels of missing ACS data in 2020) to assess its performance in simulated data scenarios, and to compare it to alternative nonresponse weighting techniques (e.g., inverse propensity weighting-IPW) and (2) produced a draft report describing the IPW approach including customized Maximum Likelihood algorithm tuning, consideration for multiple stage propensity models, comparison balance between respondent and nonrespondent characteristics, visualizations of geographic variation in model performance, and model comparisons between General Linear Models and boosting. [CSRM (Morris, Kang, Joyce, Dompreh, Slud); DSSD (Asiala, Castro); CES (Eggleston); SEHSD (Rothbaum)]
- A CSRM researcher and colleague in Research & Methodology Directorate published a journal article for addressing a "change of support" problem through the use of continuous-time models of the underlying population process, while taking into account the sampling error that comes with sample survey data. [CSRM (Joyce), R&M (McElroy)]
- CSRM researchers published a *CSRM Research Report* providing a machine learning approach and published a journal article providing a Bayesian Approach for providing counts for language minority groups in the United States in support of the *U.S. Voting Rights Act, Section 203* requirement on the Census Bureau to determine which jurisdictions must provide voting materials in languages in addition to English. [CSRM (Kang, Hall)]
- CSRM researchers, colleagues in the Demographic Statistical Methods Division, and an ASA/NSF/Census Research Fellow completed results and began drafting a research report on nonresponse weight adjustment for the former Household Pulse Survey (now Household Trends and Outlook Pulse Survey) using simulation and data analysis comparing a variety of weight trimming and smoothing procedures. [CSRM (Morris); DSMD (Chestnut, Bauder); ASA/NSF/Census Research Fellow, The University of Oklahoma Health Sciences (Chen)]
- CSRM researchers and colleagues in the Economic Statistical Methods Division, and Economic Management Division: (1) finished constructing a prior sensitivity analysis for the chosen small area models for the

re-engineered sample survey called the Annual Integrated Economic Survey (AIES); (2) obtained point estimates and design variances from the AIES; and (3) finished developing code to produce small area estimates for production. [CSRM (Aleshin-Guendel, Maples, Datta, Janicki); ESMD (Kaputa, Thompson); EMD (Madison)]

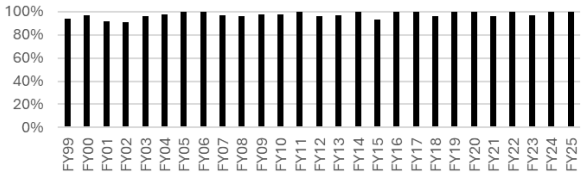
- CSRM researchers and colleagues in the Economic Statistical Methods Division, Center for Optimization and Data Science, and Research & Methodology Directorate implemented a series of updates and addressed fixes for defects from previous versions of seasonal adjustment software for the planned release of Build 62, the *X-13 ARIMA-SEATS* seasonal adjustment software in quarter four of FY2025. [CSRM (Livsey, Pang); ESMD (Lytras); CODS (Sun, Devictor); R&M (McElroy, Bell)]
- ASA/NSF/Census Research Fellow, The University of Oklahoma Health Sciences (assisted by CSRM) planned and presented (August-September, 2025) a series of 5 lectures revealing advances and knowledge on merging data from probability and nonprobability samples under the theme *Analytic Tools for Nonprobability Samples*: (1) "Introduction to Probability and Nonprobability Sampling"; (2) "Calibration Methods"; (3) "Propensity Score Methods"; (4) "Mass Imputation Methods"; and (5) "Doubly Robust and Multiply Robust Methods". [ASA/NSF/Census Research Fellow, The University of Oklahoma Health Sciences (Chen)]
- The ASA/NSF/Census Research Fellow (The University of Oklahoma Health Sciences) gained journal acceptance for research results on integrating Probability and Nonprobability samples through deep-learning-based mass imputation. [(Chen)]
- CSRM researcher published a journal article which presents and explores a framework for controlling uncertainty and tightness in a joint confidence region for an estimated ranking by optimizing the allocation of sample sizes across K populations. [CSRM (Wright)]
- CSRM program analyst/PhD in statistics candidate and an external researcher gained journal acceptance of an article with research results that (1) considers the properties of regression estimation under a misspecified sampling design, in which the nominal and true sample inclusion probabilities do not necessarily match; (2) presents an asymptotic analysis of the regression estimator, an expression of the bias, and an expression of the variance; and (3) derives a consistent variance estimator and an expression which estimates the bias in-part or in-whole. [CSRM (Engmark); University of Maryland, College Park (Opsomer)]
- CSRM and external researchers gained journal acceptance of an article for predicting census survey response rates with parsimonious additive models and structured interaction. [CSRM (Ben-David); MIT (Ibrahim, Mazumder); University of Sydney (Radchenko)]
- CSRM and R&M researchers gained journal acceptance of an article presenting Bayesian methods to improve the accuracy of differentially private measurements of constrained parameters. [CSRM (Janicki, Irimata, Livsey, Raim); R&M (Holan)]
- CSRM researcher and an external researcher published a journal article providing theoretical results for multiple imputation for parametric inference under a differentially private Laplace Mechanism. [CSRM (Sinha); FDA (Klein)]
- A CSRM researcher published a journal article presenting and discussing a review of modern multivariate -derived and partition-based record linkage methods. [CSRM (Thibaudeau)]
- CSRM and R&M researchers gained acceptance of a book chapter on the analysis of official time series with Ecce Signum, an R package for multivariate signal extraction and forecasting. [CSRM (Livsey); R&M (McElroy)]
- CSRM researcher and R&M researchers completed testing of model-imputed detailed race and ethnicity data for test datasets from eight states; using a suite of evaluation metrics, staff documented results in a draft report for Population Division, showing that bigLC software's model-based imputation performs comparably to the established Census Bureau hot-deck method, positioning it as a viable option for broader use within the Census Bureau. [CSRM (Ben-David); R&M (Mule, Schafer)]

How Did We[1] Do...

For the 27th year, we received feedback from our sponsors. Near the end of fiscal year 2025, our efforts on 28 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 28 questionnaires were obtained with the following results (The graph associated with each measure shows the performance measure over the last 27 fiscal years):

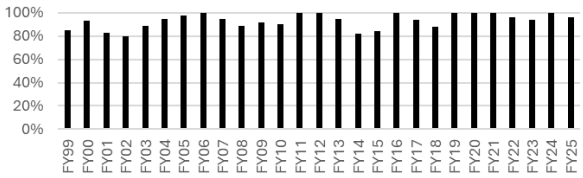
Measure 1. Overall, Work Met Expectations

Percent of FY2025 Program Sponsored Projects/Subprojects where sponsors reported that overall work met their expectations (agree or strongly agree) (28 out of 28 responses) ……..…..100%



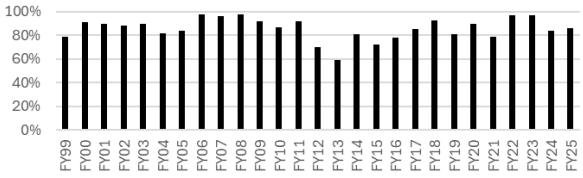
Measure 2. Established Major Deadlines Met

Percent of FY2025 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met (27 out of 28 responses) ……....…….96%



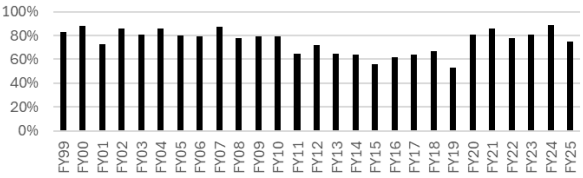
Measure 3a. At Least One Improved Method, Developed Technique, Solution, or New Insight

Percent of FY2025 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight (24 out of 28 responses) ……..86%



Measure 3b. Plans for Implementation

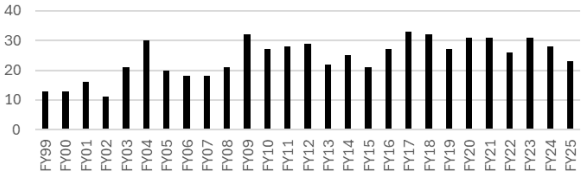
Of these FY2025 Program Sponsored Projects/Subprojects reporting at least one improved method, technique developed, solution, or new insight, the percent with plans for implementation (18 out of 24 responses) ……..75%[2]



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

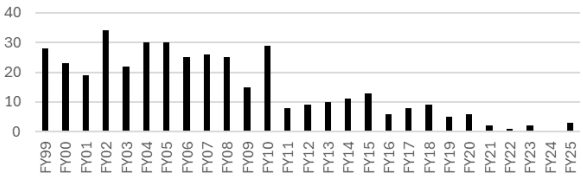
Measure 4. Journal Articles (Peer-Reviewed), Publications

Number of peer reviewed journal publications documenting research that appeared (15) or were accepted (8) in FY2025 ……23



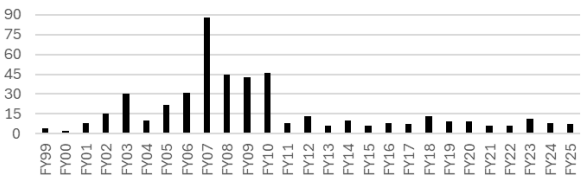
Measure 5. Proceedings, Publications

Number of proceedings publications documenting research that appeared in FY2025 ……..…..……..…3



Measure 6. Center Research Reports/Studies, Publications

Number of center research reports/studies publications documenting research that appeared in FY2025 …….7



Each complete 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.
[2]Percentages for Measure 3b previously reported in annual reports for years 2020-2024 were too low and have been corrected here.

TABLE OF CONTENTS

1. COLLABORATION

1.1 PERSON CHARACTERISTIC FRAME (Decennial Project 5350M04)

1.2 IN-OFFICE ENUMERATION (Decennial Project 5450M20)

1.3 RESPONSE QUALITY ASSURANCE FOLLOW-UP (Decennial Project 5450M21)

1.4 RESPONSE PROCESSING PLANNING & SUPPORT (Decennial Project 5450M23)

1.5 DATA PRODUCT CREATION AND DISSEMINATION (Decennial Project 5550M01)

1.6 REDISTRICTING DATA PROGRAM (Decennial Project 5550M02)

1.7 EVALUATIONS, EXPERIMENTS, AND RESEARCH (Decennial Project 5650M01)

1.8 POST ENUMERATION SURVEY DESIGN & ESTIMATION (Decennial Project 5650M02)

A. Study on Children 10-14 Years Old Counted in Census 2020 but Uncounted in the 2010 Census

Description: One of the persistent issues in the Decennial Census is undercoverage of young children. There is some evidence that much of the issue is parents simply failing to report very young children. This study plans to look at this issue by examining the status in the 2010 Decennial Census of children aged 10-14 in the 2020 Decennial Census.

Highlights: During FY 2025, staff wrote and updated a draft overview of the proposed research. Staff then wrote a draft report on overall results. The draft gives an overview of data and methodology, some results on overall nonmatching rate and the effects of age group and allocation flag and some results on consistency of Census 2020 age and Census 2010 age. Staff also wrote a draft note comparing results for Census 2020 persons with age only and Census 2020 persons with date of birth only. Staff also wrote a draft report looking at results for demographic subgroups for children aged 10-14 in 2020 and a second draft report comparing results for demographic subgroups for children aged 10-14 in 2020 and children aged 15-17 in 2020.

Staff: Michael Ikeda (x31756)

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 2025, there was no significant progress in this project.

Staff: Michael Ikeda (x31756)

C. 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 2025, staff investigated the use of statistical modeling approaches to produce population counts for the PL94-171 redistricting data product, subject to hierarchical and logical constraints. Staff evaluated the use of approaches such as iterative proportional fitting at the block and tract level for detailed race groups consisting of three or more races. Staff created a report for management review, which details the results of this work. Staff also continued to evaluate the use of regression modeling approaches for Supplemental Detailed Housing Characteristics (S-DHC) tables and compared the use of a constrained regression model with

a log-linear regression model with a many-to-many design matrix to account for covariates in the model. The results of the S-DHC modeling comparison were recorded in a research report, which includes results for one ratio and one count table and which is currently under review. During FY2025, staff also completed revisions to a manuscript detailing the use of a Bayesian model with a constrained prior for modeling S-DHC tabulations, which was published in *Journal of Privacy and Confidentiality*.

Staff: Kyle Irimata (x36465), Andrew Raim, Ryan Janicki, James Livsey, Adam Hall, Jerry Maples, Scott Holan (R&M)

D. 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 2025, staff were involved in extensive R coding and implemented the simultaneous edit-and-imputation framework using bigLC for nested household–person race and ethnicity data, including challenging substitution cases in which entire households had all items missing. Replicated comparisons of model-based latent class imputation versus hot deck under Census edit rules showed competitive performance and some gains for the bigLC approach, supporting its use for large-scale characteristic imputation. Staff extended this framework to the Population Division's Race & Ethnicity Imputation Testing Group files for testing, completing edit and imputation for test data that include 255 ($=2^8-1$) detailed combinations of eight race/ethnicity categories and continuing Stage III work for roughly 40,000 detailed categories. Staff completed testing of model-imputed detailed race and ethnicity data for test datasets from eight states (Alaska, California, Hawaii, Indiana, Maine, Montana, New York, and West Virginia); the very large California (~37 million records) and New York (~14 million records) files required specialized implementation techniques. Using a suite of evaluation metrics, staff documented results in a draft report for POP, showing that bigLC's model-based imputation performs comparably to the established Census hot-deck method, positioning it as a viable option for broader use within the Census Bureau.

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

E. 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 2025, staff continued development of panel count models and collaborated with DSSD technical staff to transition materials from research to application. Initial focus was on planning and analysis for the 2026 Census Test. Models express counts of responses over intervals of time for tracts or other geographical units, adjusting for exposure to the MQA operation and other covariates.

Staff developed three R packages as a reference implementation for modeling ideas proposed earlier in the project. "MQAPanelCountModel" implements maximum likelihood using numerical optimization as well as Bayesian fitting through Stan. "MQAProximity" provides functions to compute a spatio-temporal "proximity" metric that associates tracts with times and locations of MQA events that occurred in the vicinity; this can serve as a tract's exposure to the MQA operation. The third package "MQAModelUtilities" is a collection of utilities to support the previously mentioned packages. Packages are maintained and accessible in source control. They include comprehensive documentation: vignettes to describe packages, Roxygen to describe specific functions, and brief guides to demonstrate workflows for use cases of model-assisted MQA placement.

Staff: Andrew Raim (x37894), Lisa Moore (DCMD), Brian Roberts (DSSD)

F. 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 2025, staff ran a simple imputation method on version 4 of the 2022 Demographic File, version 4 of the 2020 Demographic File, and version 2 of the 2023 Demographic File and made the final output files (along with documentation) available to several Census Bureau researchers. Staff examined the use of IRS records given names data to assign sex for version 4 of the 2022 Demographic File. Staff also examined the use of IRS family names data to assign Hispanic Origin for both versions 1 and 4 of the 2022 Demographic File. For both sex and Hispanic Origin, staff assigned using three methods: proportions for specific names at the state level, proportions for specific names at the national level, and a combined method which assigns if either the state or national assigns unless the two assignments are contradictory. For both sex and Hispanic Origin, the three methods perform similarly although the combined method assigns to a slightly higher proportion of records. Staff also examined the relationship between IRS status and age group after Census Edited File (CEF) data replacement for version 4 of the 2022 Demographic File. 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. Staff wrote five draft reports, one summarizing the results of using IRS family names data to assign Hispanic Origin for version 1 of the 2022 Demographic File, one summarizing the imputation results for version 4 of the 2022 Demographic File, one summarizing the effect of using IRS given names data to assign sex for version 4 of the 2022 Demographic File, one summarizing the imputation results for version 4 of the 2020 Demographic File, and one summarizing the imputation results for version 2 of the 2023 Demographic File. The draft reports for the 2023 version 2 file and the 2020 version 4 file included looking at the extent to which a known bug in the creation of these two files (missing age for persons with a specific birth year) could be covered by using age from other versions of the Demographic File. It appears that most of the missing age due to this bug that cannot be obtained from the Census Edited File can be obtained from versions of the Demographic File that do not have this bug. Staff also made minor revisions to the draft report summarizing the results of using IRS given names to assign sex for version 1 of the 2022 Demographic File.

Staff: Michael Ikeda (x31756)

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 2025, staff produced estimates of *PL94-171* block group level reliability for privacy-protected population counts from the TDA. Staff identified appropriate models for producing these estimates using statistical modeling approaches. Staff showed that the predicted reliability levels in the *2020 PL94-171* TDA data were similar to the reliability levels in the 2010 demonstration data product. Staff published a *CSRM Research Report* that presented the results of this modeling exercise, as well as the predicted reliability levels for the *2020 PL94-171* TDA counts.

Staff: Kyle Irimata (x36465), Tommy Wright

H. 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 2025, staff focused on triple-system estimation (TSE) on the Demographic Frame (DF) by breaking it up into three databases: IRS1040, IRS1099, and a composite of the remaining databases not including VSGI (a commercial database). State-level estimates were then obtained that were in accord with the published estimates. Their work was published as "Triple System Estimation of National Population Counts Through Log-linear and Latent Class Modeling." Later, staff performed simulation studies to validate the accuracy of the method. They found that their estimates were unbiased in the presence of measurement error among all three databases.

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

I. 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. However, we 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 2025, staff advanced the birth cohort component study with the aim of producing birth counts by state, gender, race, and ethnicity. The primary source of birth information, the Numident data file from the Social Security Administration, generally lacks race and ethnicity information for births. The team developed an approach that uses a Demographic Frame and the Census Household Composition Key (CHCK) to impute these characteristics. To support this work, staff linked the three underlying datasets to create a combined file for building and evaluating the predictive model.

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

J. 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 2025, staff developed a multivariate model for small area estimation of domestic migration by demographic group. One year American Community Survey (ACS) data was used to create direct estimates of domestic migration, and various auxiliary data sources, including Demographic Frame, economic indicators and labor statistics were collected as potential predictors. A LASSO step was implemented to select the most relevant subset of the auxiliary data as a set of covariates. Staff fit the proposed multivariate small area model to the collected data and compared results to other estimation methods. Code was written to create an efficient workflow, to fit the models, and to perform diagnostics. An imputation step was also implemented for prediction in areas where there is no survey sample. A technical report documenting this work was written.

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

K. 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 2025, staff provided software and technical support for DSSD staff documented the name and address standardizer source code. Staff also provided support the for the *Matcher*, also known as *SRD Matcher*, in support of one-to-one matching for Decennial Support, Staff began modifications to *BigMatch* to allow AI to translate the code to python. The routine that manages the parameters of matching required most work and would be most supported by a translation to python. Furthermore, the address and name standardizers source code have been documented. Support and python translation has begun for the *SRD Matcher*. Staff presented "Graph Matching" to the Record Linkage Brownbag Seminar.

Staff: Ned Porter (x31798), Dan Weinberg

L. Cohort Component International Migration Modeling (Continuous Count Study)

Description: The focus of international migration in Continuous Count Study (CCS) Cohort Component Modeling project is to first understand the contemporary methods and challenges in measuring international migration flows and then considering improvements to the methods either via modeling exercise or through the consideration of supplementary administrative records. International migration can be related from its four main components, net migration for U.S. born persons, net migration for Puerto Rican born persons, immigration of foreign-born persons, and emigration of foreign-born person. Upon understanding the methods of estimation currently employed by the U.S. Census Bureau, one then can extend considerations of models, techniques, and data

to novel situations as they arise. The main goal of methods development is to improve estimation but also to quantify statistical uncertainty. As many of the components of international migration depend upon the American Community Survey (ACS), there are many aspects for which statistical uncertainty is immediately quantifiable.

Highlights: During FY 2025, staff completed and submitted a proceedings paper concerning the residual method of foreign-born emigration estimation. This paper served two functions: it detailed the residual emigration estimation process and foreign-born emigration procedures, including advances in variance calculations implied by American Community Survey (ACS) Public Use Microdata Sample (PUMS) data, presented across several tables; it also proposed an alternative rate estimator for foreign-born migration. Tables for estimates and variances, analogous to the PUMS data, were made available on the identical ACS vintages to staff in the Population Division (POP), the Center for Statistical Research & Methodology, and the Research & Methodology Directorate (R&M). Alternative estimation methods were investigated to improve stabilization, specifically involving the use of BigLC software, Bayesian zero-truncation, and the use of ACS data with "residence one year ago" walkbacks with experiments on sub-national estimates. Summary documentation and tables were subsequently provided to POP and R&M staff.

Staff: Patrick Joyce (x36793), Eric Slud, Tom Mule (R&M), Eric Jensen (POP), Esta Miller (POP)

M. 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 FY2025 and due to retirement of a staff member, limited progress was made and work on this project has been suspended.

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

N. 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 2025, staff regularly participated in discussions with the decennial coverage staff. The team has been investigating issues and formulating plans for how to adapt coverage measurement to using administrative records rather than a post enumeration survey. Some discussions have explored the utility of having a limited field follow-up and how to maximize that information for improving coverage measurement with the changes in the way race will be measured for the 2030 Census; staff is working on methods to make the administrative records consistent with the new racial categories.

Staff assisted on a special short-term project to see the accuracy of enumeration using only the internet self-response form and administrative records for both follow-up and non-response. This exercise used many of the ideas that have been discussed for census coverage measurement for 2030. The results from the exercise demonstrated that an enumeration was possible based on no field follow-up; however, the errors would be unacceptably high.

Staff: Jerry Maples (x32873), Ryan Janicki

O. Agreements for Advancing Record Linkage

Description: Motivated by the enhanced needs at the Census Bureau regarding the state-of-the-art methodology and algorithms for record linkage and entity resolution, four universities were awarded priority one and two cooperative agreements: the University of Arkansas, Little Rock; The University of Connecticut; The University of Michigan; and The University of Washington.

Highlights: During FY 2025, the four universities proposed developments of record linkage methodology and algorithms, including open-source software and publishing peer reviewed papers that could be of use at the Census Bureau. They engaged with staff in bi-weekly meetings to include progress reports, given presentations, and received feedback. Specifically, the University of Michigan has focused on research that will improve capabilities for record linkage that involves researchers with multiple backgrounds, including statistics, survey methodology, economics, history, computer science, and demography. The University of Washington has focused on creating data pipelines and providing simulation software based on Census Bureau data. The University of Connecticut has focused on improving computational

methods for record linkage, such as blocking, as to avoid all-to-all record comparisons and proposed a record linkage algorithm that integrates blocking with LLM-based similarity measures. Experimental evaluation showed that the approach achieves lower runtimes and higher F1-scores compared to methods that rely solely on LLMs. These findings highlight the continued importance of blocking, even in the era of advanced machine learning models. The University of Arkansas, Little Rock has focused on methods to improve the schema alignment stage of the data, cleaning pipeline, such as naming and addressing, passing using machine, learning and graph based techniques.

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

P. 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 2025 and due to loss of staff, this work has been suspended.

Staff: James Livsey (x33517), Eric Slud

Q. Cohort Component Death Modeling (Continuous Count Study)

Description. The focus of death modeling in the Continuous Count Study (CCS) Cohort Component Modeling project is to combine data sources on death provided to the U.S. Census Bureau, specifically vital record information, Numident (from the Social Security Administration) information, the various sources that produce the demographic frame, as well as yet unspecified sources of death information in order to provide a contemporary estimate of death which provides accurate information on age, sex, race, and ethnicity information to a subnational level, state at minimum and county is desired. Each data source comes with inherent flaws. Vital records from an external government agency (NCHS/CDC) have been the incumbent method of detailed resident death information but this information is lagged at one-year for totals at a national level and two-years for any further detail. The Numident provides information about known deaths in the Social Security Administration (SSA) data but the SSA retains scant information other than date of birth and date of death and the scope of the SSA is those people with Social Security Numbers (SSNs). Not all residents in the population will have an SSN and not people with an SSN will be resident within the United States. Numident is the source of death information on the demographic frame but the utility of the demographic frame is limited up to its ability to match to specific persons and their residencies and therefore is subject to coverage issues. The challenge of this project will be to find a way to propose a suitable death number (e.g., national, state, sub-state) in collaboration with the subject matter experts in the Population Division.

Highlights: During FY 2025, staff began an investigation into the sourcing of death information from the U.S. Census Bureau's various products. The complete sourcing, and the partiality thereof, shall dictate the appropriate model forms. The investigation has involved examining the demographic frame and Numident files as well as asking questions of contact points within the Census Bureau about both products. The demographic frame has so far shown significant missing data regarding key non-age demographic information. Ongoing work involves further documentation of information of the Numident and demographic frame, as well as inquiries into the availability of vital record information and the feasibility of transferring that information into a research computing environment.

Staff: Patrick Joyce (x36793), Tom Mule (R&M), Eric Jensen (POP)

1.9 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 6385M70)

A. 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 2025, staff continued regular conversations with colleagues in the Decennial Statistical Studies Division, Center for Economic Studies, and Social, Economic, and Housing Statistics Division to discuss estimate evaluation tools and criteria, variance estimation, model refinements for ACS 5-year data modeling, and innovations to the experimental Entropy Balance Weighting (EBW) method. Staff continued work with a research dataset on response status and administrative records for 2018-2022 ACS sampled units. With this dataset, staff further developed, assessed 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 updated 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). Staff produced a draft technical report describing the IPW approach including customized ML algorithm tuning, considerations for multiple stage propensity models, comparisons balance between respondent and nonrespondent characteristics, visualizations of geographic variation in model performance and model comparisons between GLM and boosting. Staff presented material from the technical report at international and national conferences.

Staff developed a Shiny app to explore various comparisons of weighting methods, geographies, variables and years. Based on propensity model evaluation metrics, logistic regression and Gradient Boosting Machine (GBM) modeling 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 by geography, variable and year 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 also worked on developing static geographic visualizations of nonresponse model comparisons to illuminate state-level variation in lack of fit.

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

B. 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 2025, no significant progress was made. However, a peer-reviewed article related to this work (Joyce and McElroy, 2024) was given a publication date and page number. Aside from this, some discussion of the topic came up in reference to evaluating monthly American Community Survey (ACS) series information in relation to collaboration 1.9.A. of this document using a time series framework.

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

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 2025, staff led the development and production-readiness of the next-generation statistical models for the Voting Rights Act (VRA) project, significantly elevating the reliability and accuracy of small area estimations. This work included overseeing the Bayesian Polya-gamma model, guiding research into prior distributions for variance reduction, and delivering a methodology paper published in the

Mathematics journal. Furthermore, the staff implemented a robust Random Forest (RF) model, validated through rigorous simulation for superior performance, which was subsequently published as a Census Bureau Research Report.

Staff also drove cross-functional efforts, coordinating with SAS programmers to integrate 2023 geographical files and updated language minority groups into the baseline *VRA* data. Staff implemented the production level of the *VRA* model for efficient testing of the 2023 ACS 5-year data.

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

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

Highlights: See reference in above *Description*.

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

1.10 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 2025, completion of a manuscript and its review were anticipated. However, due to retirement of a staff member, work on this project has been suspended.

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 (now Household Trends and Outlook 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 2025, staff worked with a team of Demographic Statistical Methods Division staff on nonresponse weight adjustment for the former Household Pulse Survey (HPS), now Household Trends and Outlook Pulse Survey (HTOPS). Staff got familiar with technical details of the new HTOPS processing including imputation and calibration. Staff studied and assessed smoothing and trimming methods for extreme weights using a variety of modeling techniques including machine learning algorithms. Even under simple sampling plans, weight adjustments during the post-processing of the survey data can result in a long-tailed weight distribution. Weight smoothing or trimming introduces estimate bias but potentially significant decreases in variance. Staff studied this trade-off empirically with historic HPS public microdata and compared and assessed variance estimation using Taylor linearization versus replication methods such as successive differences. Staff also continued applying geographically-varying visualizations of nonresponse patterns and outcome estimation bias as applied to potential nonresponse adjustment procedures for the HPS.

Staff: Darcy Steeg Morris (x33989), Sixia Chen (ASA/NSF/Census Research Fellow; The University of Oklahoma Health Sciences), John Chestnut (DSMD), Mark Bauder (DSMD)

1.11 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 2025, staff suspended work on this project and plans to finalize and send a draft report for technical review in FY 2026.

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 2025, staff implemented and analyzed a two-stage nonresponse adjustment for the CPS to improve estimation of the unemployment rate. The first stage applies a propensity score–based nonresponse adjustment, using ensemble methods (including random forests and gradient boosting) and incorporating covariates from administrative data linked to CPS. The second stage calibrates the resulting adjusted weights. For both stages, staff conducted extensive variable selection for administrative records covariates, combining Lasso (via the *glmnet* package in R) with L0 regularization (via the *L0Learn* package) to identify the most predictive variables. In addition, under the CPS Nonresponse Adjustment work, staff implemented proxy pattern-mixture models for an initial analysis in a more general setting, allowing for varying degrees of informative and nonignorable nonresponse.

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.

Highlights: During FY 2025, 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 CPS files for evaluation. Staff wrote data cleaning software for the CPS data and ACS data. Staff also wrote data cleaning software for CPS and ACS data for linking of the variables. Staff wrote programs to manage the links between administrative data and public use microdata files.

Staff: Ned Porter (x31798)

1.12 SOCIAL, ECONOMIC, & HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165025)

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 2025, staff consulted with the team members in SEHSD about the timing of an evaluation of using the Dirichlet Multinomial models for the school district poverty and population estimates. Staff continued to explore new model frameworks that would allow for correlations between two sets of shares, which would allow the shares of children in poverty and not in poverty to be dependent on each other.

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 2025, staff, with help from the SEHSD team members, are planning to apply the geographically weighted models to the one or more of the historic county-level production datasets. First, the SAIPE models will be evaluated based on standard geography (adjacency of counties) and a second approach will be based on a metric constructed from characteristics about the counties such as population size, percent minority population, percent rural population, etc.

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 2025, there was no significant progress on this project.

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

1.13 GENERAL ECONOMIC STATISTICAL SUPPORT (Economic Project 1183X01)

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

A. 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 2025, staff transitioned the target level for inference from 3-digit NAICS codes to 4-digit NAICS codes. Staff finished conducting simulation studies, including a prior sensitivity analysis, an exploration of the use of smoothing models for the design-based survey variances, and a reproduction of earlier simulations at the 4-digit NAICS level. Staff developed code to produce small area estimates for production and are now in the process of obtaining final production estimates.

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

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 2025, staff provided assistance with helpline requests from various organizations regarding the use of the X-13ARIMA-SEATS seasonal adjustment software including First National Bank of Botswana, ASL Capital Markets Inc., and Brevan Howard. Staff also provided assistance with a helpline request from India for seasonally adjusting some of their economic series.

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 2025, staff implemented a series of updates and fixes for Build 62 of *X-13ARIMA-SEATS*. The arguments appendfcst and appendbcst were added to the forecast spec. Several defects from previous versions were also addressed:

- The -c flag was modified to sum the components of a composite adjustment, but to perform modeling or seasonal adjustment only on the total.
- The critical values used in Cochran's C test when setting calendarsigma = signif were updated.
- A fatal error could occur when the graphics directory and output directory were the same, and there were save file requests within the spec file; this was fixed.
- An array value was reinitialized when running a new series to prevent a potential program crash.
- An invalid memory reference issue in the ASCII version was fixed.
- An out-of-index error when running a spec file in graphics mode was fixed.

The release also addressed several typos found in the documentation. Build 62 was released to the public at https://www.census.gov/data/software/x13as.html.

Staff: James Livsey (x33517), Demetra Lytras (ESMD), Tucker McElroy (R&M), Osbert Pang, William Bell (R&M), Lijing Sun (CODS), Antoine Devictor (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 2025, staff continued exploring the modeling of weekly data using non-integer periodicity differencing operators. This involved specifying a unit root structure for time series with fractional periodicities; this allows for convergence to standard seasonal differencing operators with integer periodicities. The weekly unit root operator was further examined to determine the position of each unit root along the unit circle, allowing users to more readily visualize the functional form applied to their time series. Staff implemented these changes in Ecce Signum (sigex), our custom structural component model codebase, available at www.github.com/tuckermcelroy/sigex.

Additionally, staff completed a paper detailing a method for mitigating residual seasonality in hierarchically adjusted time series, with a specific focus on U.S. GDP and its sub-components; this paper is scheduled to appear in the Journal of Business and Economic Statistics. Staff also resumed work on the use of weather regressors in modeling and seasonally adjusting regional housing starts data, with an eye towards shifting from a weighting scheme based on city populations to one based on county populations.

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 2025, staff implemented changes and corrections to the *X-13ARIMA-SEATS* documentation. This included addressing minor errors such as typos and formatting inconsistencies along with more substantive updates to clarify core capabilities. A corrected description of the -c flag was added to Chapter 2. We revised the descriptions of the appendbcst and appendfcst arguments, to acknowledge their application within the forecast spec and corresponding output tables in other specs. We updated the quick reference manuals to reflect the amendments to the documentation.

All of these changes were published as part of Build 62 of *X-13ARIMA-SEATS*, available at https://www.census.gov/data/software/x13as.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 2025, staff documented their findings on a project relating to a study demonstrating the potential of multivariate signal extraction and an empirical analysis of the M3 aggregates for indirect adjustments. The results of a simulation study showing that multivariate signal extraction could outperform univariate methods in some cases, and perform comparably in others, was included in this document. This project is suspended.

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 2025, there was no significant progress.

Staff: Adam Hall (x32936)

1.15 PROGRAM DIVISION OVERHEAD (Census Bureau Project 0331000)

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

Staff: Tommy Wright (x31702), Joseph Engmark, Michael Hawkins, 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 2025, staff continued working with the IRE Technical Team which supports researchers at the Census Bureau, the Bureau of Economic Analysis, and at the Federal Statistical Research Data Centers (FSRDCs) across the nation as they work on over 1,400 projects. Because a staff member retired, CSRM work has been suspended.

Staff: Chad Russell (x33215)

1.16 NATIONAL CANCER INSTITUTE (Census Bureau Project 7225157)

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 2025, staff completed work on small area estimation (SAE) research for the National Cancer Institute with specific models for 13 tobacco smoking outcomes. Staff performed SAE modeling with Arcsine transformation methods to fit Bayesian hierarchical models using Gibbs sampling techniques within MCMC to simulate successive draws from the posterior distribution for model-based estimates at the states and county levels. Staff produced county-level direct design-based and model-based estimates for all states and 3,142 counties. Staff calculated posterior probabilities via MCMC simulations.

Additionally, staff calculated the design effect estimates for thirteen smoking outcomes at the state and county levels using 2022-2023 TUS-CPS files with urban-rural county information. Staff conducted model diagnostics for all 13 tobacco use smoking outcomes.

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

2. RESEARCH

2.1 GENERAL RESEARCH AND SUPPORT (Census Bureau Project 0331000)

Missing Data & Observational Data Modeling

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

Research Problems:
• Simultaneous imputation of multiple survey variables to maintain joint properties, related to methods of evaluation of model-based imputation methods.
• Integrating editing and imputation of sample survey and census responses via multiple imputation and latent variable models.
• Nonresponse adjustment and imputation using administrative records, based on response propensity and/or multiple imputation statistical and machine learning 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 and/or alternative source data.
• Statistical techniques (e.g., classification methods, multiple imputation models) for using alternative data sources to supplement field data collection.
• Evaluation and visualization of nonresponse bias and nonresponse adjustments for geographic and social-economic subpopulations.

Current Subprojects:
• Data Editing, Imputation, and Weighting for Nonresponse. (Morris, Thibaudeau, Kang, Ben-David, Chen, Shao)
• Imputation and Weighting Models using Observational/Alternative Data Sources. (Morris, Kang, Thibaudeau, Dompreh, Joyce)

Potential Applications:
• Study flexible and data-driven nonresponse weight adjustments using administrative records for surveys experiencing data collection interruptions such as the ACS during the COVID-19 pandemic.
• Measure sensitivity of estimates, impact of nonresponse on representativeness, and weight distributions in low-response surveys such as the Household Trends and Outlook Survey (formerly Household Pulse Survey).
• Re-visit traditional missing data model techniques (e.g. imputation and response propensity models) using machine learning algorithms with alternate data sources for household surveys such as the ACS.
• Produce multiply imputed, synthetic and/or composite estimates of more geographical granular and timely economic activity based on third party data.
• Study joint multiple imputation of categorical characteristic data in the Decennial Census using models that account for household hierarchical structure and produce plausible values that do not violate edit constraints.

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 2025, 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 sample survey design. Staff is working with a cross-directorate team to synthesize, summarize and document nonresponse bias analyses for surveys across the Census Bureau with the goal of providing examples and guidelines on assessing nonresponse bias and recommending actionable steps to adjust for such findings, such as improving calibration targets or developing propensity models.

Staff continues hands-on learning of such methods through research motivated by studying the American Community Survey (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 an inverse probability weighting 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. 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. Specifically, staff is assessing estimation of variance for the American Community Survey (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 is also studying the proper implementation of weight trimming for replicate weights when estimating variance in light of extreme adjusted survey weights.

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 implementing novel latent class model fitting on simultaneous edit and imputation via Bayesian computing and worked with 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 differential subgroup coverage rates to help develop a strategy for further nonresponse adjustments via calibration variable selection and assessed the distribution and contribution of extreme weights to guide optimal tradeoffs between possible estimate bias and efficiency. Motivated by the Household Pulse Survey (HPS), staff is studying properties of various weight trimming and weight smoothing methods. Nonresponse adjustments may result in extreme weights that have an influential impact on estimated variances. The effect of these extreme weights can be mitigated via methods reduce variance at the expense of introducing bias which can be assessed empirically. These studies or weight trimming/smoothing are a post-hoc adjustment in the process of adjusting weights for nonresponse, but it is important to understand its interplay with the entire method for accounting for nonresponse bias.

Staff: Darcy Morris (x33989), Yves Thibaudeau, Joseph Kang, Emanuel Ben-David, Sixia Chen (ASA/NSF/Census Research Fellow/The University of Oklahoma Health Sciences), Jun Shao

B. Imputation and Weighting Models 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 2025, staff assessed and studied 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 the American Community Survey (ACS) estimates using machine learning Inverse Probability Weighting (IPW) nonresponse adjustments with replicate weights at the modeling level as opposed to replication weighting factors at an aggregated level.

Staff continues 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.

As part of the cross-directorate team assessing nonresponse bias analyses across the Census Bureau, staff are working towards understanding the benefits and limitations of general recommendations for the use of auxiliary data to address nonresponse in Census Bureau surveys; whether as an additional check with respect to varying response rates across administrative record variables, or to be used in adapting nonresponse bias adjustments in survey production. This knowledge-sharing exercise aims to document common experiences and guide uses of auxiliary data for nonresponse bias analysis in household and economic surveys Census Bureau-wide.

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

Record Linkage & Machine Learning

Motivation: Record linkage continues to grow in importance as a fundamental activity in statistical agencies. The number of available administrative lists and commercial files has grown exponentially and present statistical agencies with opportunities to accumulate information through record-linkage to support the production of official statistics. In addition to cost, new obstacles to traditional data collection have emerged in the form of possibly recurrent pandemics. These circumstances further motivate the accumulation of information by linking public, private and administrative files.

The solutions developed at the Census Bureau, such as *BigMatch*, have shown to perform better than competitors in general (see for example Arthun, Gilary, McGinnis, and Zamora 2025) and are highly flexible. The challenge forward is to take advantage of the ever increasing computational power made available and to expand the latest scientific advances in an equally functional set-up.

Thibaudeau (2020) describes the strides the Census Bureau and the Statistical Research Division (now Center for Statistical Research & Methodology) have made over the years. While this is impressive, more is needed. The Census Bureau must remain abreast of the ever improving state-of-the-art in record linkage and be prepared to champion its own methodologies as some of the best in the world. Our goal is to achieve the synergy of methods and software that will benefit most the Census Bureau and its mission. System portability is also an objective. The Census Bureau should have the freedom to upgrade its IT infrastructure knowing record-linkage applications will remain functional.

Research Problems:

• Multiple evaluations at the Census Bureau (see for example Arthun, Gilary, McGinnis, and Zamora 2025) have shown the record-linkage software developed in CSRM/SRD, such as *BigMatch*, perform better than open-source competitors in general. The challenge is to maintain the same versatility as the methodology of *BigMatch* is improved. An important effort in that direction was initiated to perform "Multi-file Simultaneous Record Linkage." Sadinle and Fienberg (2013) introduce a formal theory for multi-file record linkage based on comprehensive partitioning. Partitioning accounts for all possible configurations of simultaneous record matches within a set of files, thereby ensuring pairwise transitivity and preventing logical contradictions. This approach transcends traditional attempts at linking multiple files at the Census Bureau and other institutions. Those attempts were mostly based on linking pairs of records in isolation and enforcing business rules to combine the pairs and obtain multi-record matches. Muti-file record linkage, as proposed by Sadinle and Fienberg proceeds, from information on the constructs underlying all logically possible assignment of multi-record matches as a whole. As such, rigorous multi-file record linkage is a "np-hard" problem. The work of Fienberg and Sadinle (2013) and more recent work (Steorts 2015, Marchant et al. 2021, 2023) aim at finessing the computational difficulty of multi-file record linkage through probabilistic algorithms. The prospect is logically valid multi-file record linkage, which cannot be done using traditional methods (Fellegi-Sunter). This raises the potential of retrieving a full spectrum of logically valid record matches between the records of a Census, post-enumeration survey and internal or third-party administrative files simultaneously, rather than piecing together initially independent record pairs, as is mostly done at this time.

• Markov Chains Monte-Carlo (MCMC), like that powered by d-blink, give full probabilistic characterizations of the record- linkage process and are becoming indispensable for full comprehension of a record linkage process. At the same time MCMCs can be tweaked to deliver fast snapshots of the linked population. Research in that direction is crucial. Old-School programs like *BigMatch* have been greatly optimized for fast linking but lack in nuance. They need to be garnished by richer comparison schemes, such as dictionary-assisted fuzzy string comparisons.

• MCMCs and dynamic processes exclusively offer a full probabilistic characterization of record-linkage but struggle to achieve the scalability of Fellegi-Sunter and other clustering algorithm, such as latent-class analysis. Approximations improving the scalability include large-sample theory approximations and variational approximations. These approximations are known to be accurate and computationally frugal. There are also possible hybrids and compromises that are possible between MCMCs and the static approach. One is dimensionally collapsible models as described in Weinberg & Thibaudeau (2025). Models derived from algebraic geometry can be expanded "on the spot" to

reflect the dimensionality of the clusters subject to matching (entity resolution. d-blink takes care of dimensionally extending or collapsing structures automatically at the unit level, which is computationally expansive. The approach of W-T proposes fitting several dimensionally collapsed models to a specific situation. This approach offers an advantageous middle-ground if the number of models to be fitted is not too large. As the number of dimensions -matching fields- and the number of possible dimension collapse increasing it also becomes computationally onerous. Identifying the most practical solution in specific situations is the basic challenge of record linkage going forward.

- New data structure for record-linkage of multiple large lists needs to be explored. *d-blink* is an example of a more efficient data structure: Node-connected structures minimize the number of comparisons, as opposed to a traditional all pairwise comparisons. Other structures are possible, such as cyclical linked lists (Thibaudeau 1992), and should be researched.
- As new techniques continue to be implemented and experimented on various existing software (*R*, *Python*, *C*) and hardware (Windows, OSX, IRE, CAES) platforms, the dominant paradigms are emerging and work toward integration and unification, while maintaining versatility, is moving in high gear.

Current Subprojects:

- Adjusting the Statistical Analysis on Integrated Data. (Ben-David)
- Entity Resolution and Merging Noisy Databases. (Steorts, Brown/CES, Blalock/DSMD, Thibaudeau)
- Record-Linkage Support for the Decennial Census. (Ben-David, Weinberg, Brown/CES, Thibaudeau)

Potential Applications:

- Possible massive concurrent record-linkage implementations for Census 2030. The objective is counting all distinguishable persons in linked and unduplicated administrative and commercial person-level lists.
- Unduplication and record-linkage for frame construction in the demographic and economic areas.
- Re-identification through record-linking for proofing confidentiality of data lists.
- Analysis and estimation based on linked lists.
- Linking probabilistic design-based surveys to large non-probability lists and sample for probabilistic calibration.

A. Adjusting the Statistical Analysis on Integrated 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 2025, staff conducted simulation studies showing that missed matches in record linkage can substantially bias post-linkage estimates, distorting key measures used in analysis. In response, they developed and evaluated pattern mixture model–based adjustment methods that explicitly account for missed linkage patterns. Initial simulation results indicate that these methods are effective at reducing bias and have broad applicability to linked administrative and survey data, thereby improving the reliability of analyses based on linked governmental and research datasets.

Staff: Emanuel Ben-David (x37275), Guoqing Diao (GWU), Priyanjali Bukke (GMU), Martin Slawski (UVA), 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 2025, staff continued to edit a research report and rework the associated R software. Staff presented their work at the Government Advances in Statistical Programming (GASP) virtual conference with the title "Parameter Estimation in Record Linkage Through Dimensional Reduction." They released their work as a *CSRM Research Report* entitled "Algebraic Dimensional Reduction for the Fellegi-Sunter Model in Record Linkage and General Parameter Input Specification for *BigMatch*." They have applied their methodology to realistic simulated data to deduplicate college students and children who are listed at multiple addresses. A journal article on this work is in progress.

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: This project ended in FY2024.

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: This project ended in FY2024.

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 Approximations to Bipartite Record Linkage
Description: The purpose of this project is to scale Bayesian bipartite record linkage to hundreds of thousands of records in minutes, providing an alternative approach to Markov chain Monte Carlo sampling, which is both computationally demanding and is known to have convergences issues.

Highlights: During FY 2025, staff revised a manuscript with two variational approximations for bipartite record linkage, providing comparisons to the prior literature, and have submitted it for publication. Staff have also tried to identify internal data for testing purposes, however, have faced many obstacles regarding this, despite utilizing the same dataset for prior publication. Staff have utilized other datasets in the meantime due to these obstacles to move forward with a paper. Furthermore, staff has extended this approach to a context, known as streaming data, and is working to prepare this for a separate paper submission.

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

F. Generalized Microclustering Models for Graphical Record Linkage
Description: In this project, we aim to extend recent Bayesian graphical models for entity resolution by: (i) proposing a generalized model for categorical and textual data, (ii) offering guidance for practitioners through comprehensive simulation studies, and (iii) releasing open-source software for reproducibility and applied use. To that end, we construct realistic simulations relevant to the survey methodology community, generating synthetic household populations using attributes such as names, birth years, and zip codes—allowing controlled experimentation in the absence of publicly available ground truth due to privacy constraints. To our knowledge, no such systematic evaluation has previously been conducted for this class of models.

Highlights: During FY 2025, staff proposed a general Bayesian framework that satisfies the microclustering property generalizing the existing literature and bridging it together. Staff's contributions were evaluated through comprehensive series of simulation studies, benchmarking performance against recently proposed methods in the literature. We focused on the role of the prior on the linkage structure (i.e., the prior on links versus non-links) under varying levels of duplication and distortion in the data. We found that when duplication rates are low, all priors exhibit reduced precision—tending to over-estimate the number of duplicate records—with the uniform prior performing the worst. In settings with moderate to high duplication or distortion, differences among priors diminish, and their performance becomes more comparable. We find that when users expect similar levels of duplication and distortion in their data, our results can inform the selection of an appropriate combination of modeling components to improve entity resolution accuracy. This work has been submitted for publication.

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

Small Area Estimation

Motivation: Small area estimation is important in light of a continual demand by data users for finer geographic and demographic detail of published statistics and for various subpopulations. Traditional demographic and economic sample surveys designed for national estimates do not provide large enough samples to produce reliable direct estimates for smaller areas such as counties and even most states. The use of valid statistical models, along with the availability of suitable auxiliary data, can provide small area estimates with greater precision; however, bias due to an incorrect model or failure to account for informative sampling can result.

Research Problems:
• Development of models that combine data across multiple sample surveys or combines survey and observational data (non-probability samples) to improve survey estimates.
• Development of model diagnostic and model comparison tools for small area models.
• Development of small area share models for subareas

estimates (e.g., school districts or tracts).
• Development of temporal small area estimation techniques.
• Development of spatial small area estimation techniques.
• Development of more robust estimates of mean squared error of prediction by incorporating Bayesian and bootstrap methods.
• Development of area-level models to jointly estimate the survey mean and variance.
• Development of models combining both small geographic areas crossed with small demographic subgroups.

Current Subprojects:
• Bootstrap Mean Squared Error Estimation for Small Area Means under Non-normal Random Effects. (Datta, Irimata, Maples)
• Developing correlated small area share models to create estimates of school district child poverty and population. (Maples)
• Developing geographically weighted methods to assess the assumption of constant parameter values across all domains. (Maples, Dompreh)
• Assessment of mean squared errors of empirical best linear unbiased predictors for misspecified models. (Datta)
• Development of tract by demographic population estimates for non-census years using census, ACS and Demographic Frame data. (Maples, Mule/R&M, Basel/SEHSD, Ikeda, Holan/R&M)
• Development of small area models for establishment surveys for employment and receipts. (Aleshin-Guendel, Maples, Datta, Janicki, Kaputa/ESMD, Maison/EMD)
• Variance Estimation and Modeling for Privacy-Protected Redistricting Data. (Irimata)
• Bayesian Hierarchical Spatial Models for Small Area estimation (Datta, Janicki, Maples)
• Construction of Joint Credible Set of Ranks of small Area means (Datta, Maples)

Potential Applications:
• Model diagnostic and comparison tools can be applied in any small area application, from SAIPE to SAHIE, to small area models applied to SIPP, AHS, etc.
• Temporal extensions of small area models will be potentially useful for population estimates in sub-county areas in non-census years.
• Small area share models may be a replacement to the current for the current school district estimates procedures for SAIPE.
• Spatial small area models can improve estimates and provide limited disclosure avoidance for some of the ACS special tabulations.
• Small area models to estimate employment and receipts using data from the new AIES (Annual Integrated Economic Survey) at the state by NAICS-3 level.
• Joint area-level models can be used to produce estimates of the population counts, as well as the variance in the TopDown Algorithm (TDA) due to differentially private noise addition and post-processing in the PL94-171 redistricting data.

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 2025, staff worked on rigorous mathematical treatment of the estimators of model parameters of the non-normal Fay-Herriot model. Staff considered estimation of the model parameters solving a set of estimating equations. Consistency and rate of convergence results obtained for the estimating-equation-based estimators are very general, encompassing most existing and published results on this topic. It would be a major contribution to the EBLUP literature based on Fay-Herriot model.

Substantial research progress the staff made on this topic over time showed overall superiority of the developed method over the existing methods, dealing with MSE estimation of the EBLUPs of small area means for non-normal models. Staff plan to split the voluminous results obtained so far into two manuscripts for publication in appropriate journals.

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

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 2025, staff modified and adapted the work published in two papers by Chung and Datta (2022, *Survey Methodology, Vol, 48-2, pp. 463-489)* and Li et al. (2024, *Statistics and Applications, Vol. 24-2, pp. 1-21)* on HB estimation of small area means based on spatial generalization of standard Fay-Herriot model. These two papers deal with area-level SAE model based on direct estimates or some aggregated estimates for a combined set of areas. In the ongoing project, the staff considered unit-level SAE models when unit-level data are available, which is often the case in many Census Bureau SAE applications. Additionally, sampling weights for the sampled units are also available. Use of standard unit-level models ignoring these weights is inappropriate. Moreover, random small area effects often show spatial dependence among themselves. In this project, staff is using a hierarchical Bayesian estimation of SA means based on weighted spatial random effects model. Staff used PUMA dataset to fit various spatial models. Staff carried out simulations to examine gains by the proposed estimators over the existing unit-level HB estimates of small area means. Work on a manuscript reporting these results will resume shortly.

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

C. 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 2025, staff developed a Bayesian method for ranking entities or multiple populations. In the context of Census Bureau, these subpopulations are small areas. Inference for overall ranking vector by appropriately addressing the uncertainty associated with the point estimates of the ranks is important. Two substantive frequentist solutions to this problem are available. Staff is working on several manuscripts on Bayesian approach to the ranking problem. A manuscript reporting the Bayesian methodology is currently under revision for a major journal. Data analysis, simulations and theoretical developments show that the Bayesian method has several advantages over the two frequentist solutions.

The proposed Bayesian method for inference for the rank vector is very general and can be applied even to non-normal small area estimation problems. Small areas based on estimation of proportions based on a hierarchical Bayes probit regression model have been ranked based on proposed Bayesian methodology. The proposed Bayesian method has facilitated developing a ranking of the small areas in the presence of benchmarking. A reasonable solution is based on the benchmark-satisfying perturbed posterior distribution. Staff is investigating the new methodology to income data from 72 PUMA in Georgia.

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)

D. Developing Correlated Small Area Share Models to Create Estimates of School District Child Poverty and Population

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" population and poverty estimates for school districts. The methods should also produce realistic measures of the accuracy of the estimates (standard errors). The focus on this project is to extend the Dirichlet-Multinomial share models to allow for correlated outcomes. For the school district estimates, the two sets of shares that would be considered to be correlated are the school district to county share of children in poverty and the share of children not in poverty.

Highlights: During FY 2025, staff reviewed existing literature for correlated Dirichlet random variables, but none of the existing methods fit the application to school district poverty and population shares. A new approach based on a joint model for total population and subcomponents (in poverty and not in poverty) is being researched.

Staff: Jerry Maples (x32873)

E. Developing Geographically Weighted Methods to Assess the Assumption of Constant Parameter Values Across All Domains

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 2025, staff explored the use of the Geographically Weighted Small Area model on a different dataset to better understand how to generalize interpretation of the variability of the parameters. Staff is also continuing to implement a variant of the model where only a subset of the parameters are allowed to vary while the others are fixed across all areas.

Staff: Jerry Maples (x32873), Isaac Dompreh

F. Development of Tract by Demographic Population Estimates for Non-Census Years Using Census, ACS, and Demographic Frame Data
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.

This project focuses on the alternative estimates.

Highlights: During FY 2025, staff derived a method to approximate the uncertainty of the county population estimates from the Population Estimates Program using a random walk model for the 2020 data series projection that did not use the Census 2020, and compared that to the actual Census 2020 population totals for counties.

This result allowed staff to create total population estimates and measures of uncertainty for all tracts in the announced 2026 Census Test Sites using 2022 datasets including: 2022 5-year ACS, 2020 Census and 2022 Demographic Frame datasets. These results were used for comparison and evaluation with the administrative record enumerations.

Staff updated the 2022 tract estimates with the equivalent 2023 datasets. Staff have also acquired access to the 2023 MAFX, but so far none of the models have shown any additional predictive value using variables from this database.

Staff: Jerry Maples (x32873), Tom Mule (R&M), Wes Basel (SEHSD), Michael Ikeda, Scott Holan (R&M)

G. Development of Small Area Models for Establishment Surveys for Employment and Receipts
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 2025, staff wrote an article describing the methodology of the rejection sampler being used in the rejectFH R package. The rejection sampler provides a simple method to draw exact samples from the Fay-Herriot posterior, as an alternative to non-exact methods such as Markov chain Monte Carlo. Computational bottlenecks in the sampler have been sped up using eigendecompositions, where after preprocessing samples can be drawn in time linear in the number of small areas. Staff plan to submit the article to a journal in FY2026. Staff have begun research into novel multivariate spatial small area models to improve upon the small area modeling for future AIES cycles.

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

H. 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 2025, staff investigated the use of a joint SAE model to jointly model the population counts and variance in the *2010 PL94-171* demonstration data product. Staff compared various sampling approaches, including an independent Metropolis Hastings (IMH) within Gibbs approach, as well as a Vertical Weighted Strips (VWS) within Gibbs approach and showed that the VWS approach produced better estimates with better draws. Staff also investigated appropriate covariates for use in the variance submodel and identified relevant predictors which improved estimates of the sampling variance.

Staff: Kyle Irimata (x36465)

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.

Current Subprojects:
• Developing Bayesian pseudolikelihood models for unit-level data obtained from a complex sample survey which incorporate spatio-temporal dependencies. (Holan, Janicki)
• Development of change-of-support methodology for inference on regions with no direct measurement, based on observations on a distinct geographic region or grid. (Holan, Janicki, Lahiri)
• Incorporation of spatially-correlated random effects in small area models. (Aleshin-Guendel, Datta, Janicki, Maples)
• Integration of deep learning, machine learning, and model selection, with spatial modeling. (Holan, Janicki)

Potential Applications:
• Production of "gridded" data products which correspond to a regular lattice which remains constant over time.
• Improved precision and interpretability of privacy-protected decennial census tables.
• 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. Developing Bayesian Pseudolikelihood Models for Unit-Level Data Obtained from a Complex Sample Survey Which Incorporate Spatio-Temporal Dependencies

Description: Sample survey data is often clustered by demographic characteristic and geographic region. Accounting for these dependencies in a model-based setting can be useful for understanding spatial patterns, improving the precision of estimates, and making predictions where no data is available. Typically, modeling is done on aggregated survey data. Directly modeling the unit-level survey data could provide improvements as aggregation of survey data could lead to loss of information. Some of the challenges with this approach are difficulties in computation due to high dimensional data, explicitly modeling spatio-temporal dependencies, and incorporating the survey design. This research intends to address these challenges.

Highlights: During FY 2025, staff worked on developing statistical models for unit level longitudinal data obtained from an informative sample survey. The proposed model utilizes an asymmetric Laplace likelihood and incorporates structured random effects to allow borrowing of strength over space and time. The informative survey design was accounted for by exponentially weighting the likelihood using survey weights. A sampling algorithm for fitting the model which uses a data augmentation strategy was written and coded. A data example using household wealth obtained from Survey of Income and Program Participation (SIPP) public use data as a response variable and an empirical simulation study were conducted to verify the effectiveness of the proposed methodology.

Staff: Ryan Janicki (x35725), Daniel Vedensky (University of Missouri), Scott Holan (R&M)

B. Development of Change-of-Support Methodology for Inference on Regions with No Direct Measurement, Based on Observations on a Distinct Geographic Region or Grid

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. This research is 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.

Highlights: During FY 2025, staff developed statistical downscaling methods for prediction on a regular lattice using aggregated response data from irregularly shaped geographies such as tracts or block groups. The modeling framework assumes a latent spatial Gaussian random field that is observable at an aggregated level. We further assume access to a large number of potential predictors, many of which may be spurious. The LASSO was used to select the best subset from the full set of potential predictors, and the aggregated response data was regressed onto the selected covariates. The resulting fitted model can then be used to predict at the target support. Importantly, it was discovered that constant terms must be omitted from the final model due to the different resolutions of the source and target support.

Refinements of the methodology which account for spatial dependencies were investigated. Empirical best linear unbiased prediction methods using ordinary least squares residuals were implemented to adjust predictions. Method of moments and likelihood-based methods for estimation of variance parameters were developed. A simulation study based on 2020 decennial census data was conducted to evaluate the effectiveness of the proposed methodology was performed, with results indicating accurate prediction at the target support. Staff continued development of code and documentation of work.

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

C. Incorporation of Spatially-Correlated Random Effects in Small Area Models

Description: Sample surveys can provide high-quality direct estimates in regions with large sample size. In regions with small samples, small area models have proven to be useful to improve direct survey estimates by "borrowing strength" through linking geographic regions and incorporating auxiliary covariate information. The effectiveness of small area estimation methods depends critically on the quality of the covariate information. In situations where high-quality covariates are not available, spatial correlations can be leveraged to improve predictions. This project seeks to understand how to most effectively extend classical unit-level and area-level small area estimation models to spatio-temporal settings.

Highlights: During FY 2025, staff studied the importance of inclusion of spatially correlated random effects in classical unit level small area estimation models when important predictors are missing from the model. An extensive simulation study was conducted wherein the mean structure of response variables were dependent on a set of covariates, but some subsets of the covariates were not used in the model fitting process. Various spatial random effects models were compared. The sensitivity of the predictions to inclusion and exclusion of survey weights in the model fitting procedure were also investigated.

A data example was done, using income by race and gender from ACS public use micro data, with age used as a predictor. It was found that inclusion of spatially-correlated random effects was important for reducing mean squared error of model-based predictions from the direct estimates, particularly for minority groups and women. Use of survey weights helped account for the survey design and reduced bias.

Staff: Ryan Janicki (x35725), Serge Aleshin-Guendel, Gauri Datta, Jerry Maples

D. Integration of Deep Learning, Machine Learning, and Model Selection, with 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 2025, motivated by the problem of producing supplemental detailed housing characteristics tables, staff conducted exploratory work on variable selection techniques and estimation of the functional form of regression relationships when the parameter space of statistical models is constrained. Staff created a data set from auxiliary data sources such as previous decennial census tables, noisy measurements from Supplemental Detailed Housing Characteristics (DHC) and PL tables, and ACS tables for use as predictors. A literature review on neural networks with spatial data was started, and some of these known techniques were tested on the data.

Staff also developed a novel model for differentially private measurements of constrained parameters which incorporates parameter constraints into a truncated multivariate Gaussian prior distribution and utilizes covariate data for prediction. A sampling algorithm based on the work of Chen and Shao (1998) was developed. The work involves selecting an appropriate mixing distribution and a pseudo posterior distribution which approximates the analytically intractable true posterior distribution which has constrained parameters. An R package containing code for fitting this model was

written.

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

Sampling Estimation & Survey Inference

Motivation: Survey sampling helps the Census Bureau provide timely and cost efficient estimates of population characteristics. Sampling methodology remains at the center of innovation at the Census Bureau as evidenced by three recent major efforts: the Household Trends and Outlook Pulse Survey (formerly Household Pulse Survey); the Small Business Pulse Survey; and the Annual Integrated Economic Survey. Demographic sample surveys estimate characteristics of people or households such as employment, income, poverty, health, insurance coverage, educational attainment, or crime victimization. Economic sample surveys estimate characteristics of businesses such as payroll, number of employees, production, sales, revenue, or inventory. Survey sampling helps the Census Bureau assess the quality and coverage of each decennial census. Estimates are produced by use of design-based estimation techniques or model-based estimation techniques. Methods and topics across the three program areas (Demographic, Economic, and Decennial) include: sample design, estimation and use of auxiliary information (e.g., sampling frame and administrative records), weighting methodology, adjustments for non-response, proper use of population estimates as weighting controls, variance estimation, effects of imputation on variances, coverage measurement sampling and estimation, coverage measurement evaluation, evaluation of census operations, uses of administrative records in census operations, improvement in census processing, and analyses that aid in increasing census response.

Research Problems:
• How to design and analyze sample surveys from "frames" determined by non-probabilistically sampled observational data to achieve representative population coverage. To make census data products based jointly on administrative and survey data fully representative of the general population, as our current surveys are, new sampling designs and analysis methods will have to be developed.
• How can inclusion in observational or administrative lists be modeled jointly with indicator and mode of survey response, so that traditional survey methods can be extended to merged survey and non-survey data?
• 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 formulated and solved as optimization problems to avoid the ambiguities resulting from multiple weighting step and to explicitly allow inexact calibration?
• What models can aid in assessing the combined effect of all the sources of sampling and nonsampling error, including frame coverage errors and measurement errors, on sample survey estimates?
• What experiments and analyses can inform the development of outreach methods to enhance census response?
• Can unduplication and matching errors be accounted for in modeling frame coverage in censuses and sample surveys?
• How can small-area or other model-based methods be used to improve interval estimates in sample surveys, to design survey collection methods with lowered costs, or to improve Census Bureau imputation methods?
• Can classical methods in nonparametrics (e.g., using ranks) improve estimates from sample surveys?
• How can we measure and present uncertainty in rankings of units based on sample survey estimates?
• Can data from sources other than censuses and sample surveys be used to improve results from censuses and sample surveys?
• How to develop and use bootstrap methods for expressing uncertainty in estimates from probability sampling?

Current Subprojects:
• Integration of Data from Probability and Nonprobability Samples. (Wright, Chen, Mulry, Ikeda)
• The Ranking Project: Methodology Development and Evaluation. (Wright, Yau, Wieczorek/Colby College, Hall)
• Optimal Allocation Methods: Sample Allocation and Apportionment. (Wright)
• Replication Methods for Variance Estimation: Understanding Successive Difference Replication and Bootstrapping. (Joyce, Wright)

Potential Applications:
• Improve estimates and reduce costs for household sample surveys by introducing new design and estimation methods, possibly to compensate for smaller sample sizes.
• 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. Employ administrative records and other data sources to improve the estimates from probability samples.
• Measure and report uncertainty in estimated rankings in household and economic sample surveys.
• Develop bootstrap methods for expressing uncertainty as an alternative source of published variance estimates and as a check on existing methods of producing variances in Census Bureau sample surveys.

A. Integration of Data from Probability and

Nonprobability Samples
Description: With decreasing response rates to government probability-based sample surveys (e.g., Current Population, American Community Survey), this project seeks to investigate whether or not some compensation can be realized by integrating data from probability samples with the increasing sources of data (e.g., administrative data) not based on probability, i.e., nonprobability samples. Probability provides a framework for making uncertainty statements about estimates from probability sample surveys. We aim to consider the possibility of decreasing targeted sample sizes to better focus nonincreasing resources for data collection.

Highlights: During FY 2025, staff worked with the ASA/NSF/Census Research Fellow who developed and presented for all Census Bureau staff a series of lectures revealing advances and knowledge on merging data from probability and nonprobability sampling as follows (each at 11:00am): August 19 – "Introduction to Probability and Nonprobability Sampling" (Lecture 1 of 5 Lectures on Analytic Tools for Non-probability Samples); August 26 – "Calibration Methods" (Lecture 2 of 5); September 9 – "Propensity Score Methods" (Lecture 3 of 5); September 16 – "Mass Imputation Methods" (Lecture 4 of 5); September 23 – "Doubly Robust and Multiply Robust Methods" (Lecture 5 of 5).

Staff: Tommy Wright (x31702), Sixia Chen (ASA/NSF/Census Research Fellow/University of Oklahoma Health Sciences), Joseph Engmark, Mary Mulry, Michael Ikeda

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

Highlights: During FY 2025, staff added 1-Year American Community Survey (ACS) data for 2023 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 89 different American ACS topics. The 2 updated visualizations now provide ACS data for the years 2018, 2019, 2021, 2022, and 2023 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 published a paper "Optimal Tightening of the KWW Joint Confidence Region for a Ranking" which explores and provides a framework for controlling uncertainty and tightness in an estimated ranking by optimizing the allocation of sample sizes across K Populations. Staff also presented a Center for Statistical Research & Methodology (CSRM) seminar "Uncertainty Reduction for an Estimated Ranking Using Differences" which: provided theory and computation details of two methods (INDI and DIFF), noted improvements with DIFF, and shared examples from the visual Estimated Rankings of All States.

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

C. Optimal Allocation Methods: Sample Allocation 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 H 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/apportionment allocation algorithms.

Sample Allocation
Highlights: During FY 2025, staff published a paper "Optimal Tightening of the KWW Joint Confidence Region for a Ranking" which explores a framework for controlling uncertainty and tightness in an estimated ranking by optimizing the allocation of sample sizes across K Populations.

Staff: Tommy Wright (x31702)

Apportionment
Highlights: During FY 2025, staff continued: (1) investigation of a framework for determining an optimal size of the U.S. House of Representatives which uses a convenient objective function that sequentially minimizes the differences between the state ratios (where a state ratio is its population divided by its number of allocated seats) each time the house size is increased by one seat, and (2) investigation of a mathematical framework for generalizing the Method of Equal of Proportions, which allows flexibility on the way the number of seats are allocated. Research results were shared in an invited talk at the 2025 Joint Mathematics Meeting. Staff continued this investigation and began work to convert the slides from the talk into a *CSRM Research Report*.

Staff: Tommy Wright (x31702)

D. Replication Methods for Variance Estimation: Understanding Successive Difference Replication and Bootstrapping

Description: Partly due to complexity of sampling and estimation methodology, both the Current Population Survey and American Community Survey make use of successive difference replication (Wolter, 1984,2007; Fay & Train, 1995) to provide estimates of sampling variance of various estimates. For a sample of n units, the successive difference replication method produces k sets of n sample weights each and produces k different estimates. With k estimates, sampling variance of the original estimate is computed by computing the variance of the k estimates. Alternately, the bootstrap (Efron, 1979; Rao, 1984) provides a framework for estimating sampling variance of an estimate by "…resampling with replacement from the original sample, creating multiple *bootstrapped* samples and calculating the estimate of interest (like a mean) for each sample, then using the variability of these bootstrap estimates to estimate the sampling variance of the original estimate. This project seeks to understand and compare both methods of sampling variance estimation empirically and theoretically.

Highlights: During FY 2025, staff conducted a review of the literature on this topic and conducted preliminary discussions concerning the development of a project plan. A series of presentation notes began development for the self-education of immediate staff, with the longer goal of informing the Center for Statistical Research & Methodology (CSRM) and U.S. Census Bureau staff at large. These notes provide an informative background summary on replicate variance methods. The literature review progressed from initial study of half-sample methods through to the methods of Fay and Train and some discussion of generalization of Fay and Train. It is observed that the purpose of variance-preserving contrast methods is to then use these linear replicates in many non-linear or raked-based applications to satisfy other concerns. Interim documentation discusses the notions in which these contrast methods are intended to fit the perception of replication. Ongoing work concentrates on establishing a series of results for the generalizations of Fay and Train and developing results for infinitely small (limiting) perturbations implied by Fay's linear quadratic forms for variance estimators of non-linear functions. This includes their application to non-linear differentiable functions of estimators (reproducing the delta method of variance approximations) and their application to raking and calibration forms; the derivations and results for raking and calibration are in progress. Additionally, a step-by-step higher mathematical walkthrough describing the current Fay and Train methodology was written.

Staff: Patrick Joyce (x36793), Tommy Wright

E. 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 2025, staff modified an earlier definition of "tightness" where the new measure of tightness T* assumes values between 0 and 1. If T* = 0, there is no tightness. If T* = 1, there is complete tightness; the joint confidence region only contains the estimated ranking; and we have strong evidence the estimated ranking is the true ranking. *S*taff also presented a Center for Statistical Research & Methodology (CSRM) Seminar "Uncertainty Reduction for an Estimated Ranking Using Differences" which showed conditions for obtaining tighter joint confidence regions. The seminar shared the theoretically based visual which shows four things at once: (1) joint confidence region revealing uncertainty in the estimated ranking; (2) possible true rankings, beyond the estimated ranking, (3) a marginal confidence set for population k true rank; and (4) a marginal confidence set for each rank r.

Staff: Tommy Wright (x31702)

Time Series & Seasonal Adjustment

Motivation: Seasonal adjustment is vital to the effective presentation of data collected from monthly and quarterly economic sample 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.

Time series modeling and seasonal adjustment go hand in hand. Not only are time series models used for seasonal adjustment to forecast and extend the series, allowing the use of symmetric filters, but many model-based diagnostics are also used when understanding time series features. Additionally, the Economic Directorate releases thousands of time series values each month and quarter. This inevitably produces many unique challenges whenever there is a change or disruption in the sampling design or the overall economy. It is vital to stay up-to-date with flexible time series modeling frameworks to address these situations.

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.

• Diagnostics of seasonality must address differing sampling frequencies (monthly versus quarterly) and multiple forms of seasonality (cycles of annual versus weekly period), and must distinguish between raw and seasonally adjusted data.

• Multivariate modeling can not only provide increased precision of seasonal adjustments, but can also assist with series that have a low signal content. Moreover, multivariate techniques expand the class of univariate models, allowing the modeling of seasonal heteroscedasticity. This motivates the need to develop a viable multivariate seasonal adjustment methodology that can handle modeling, fitting, and seasonal adjustment of a large number of series.

• Time series data are being measured at higher sampling rates or over geographical regions, requiring new seasonal adjustment methods for high frequency/space-time data.

• Many published time series arise from sample surveys, and are subject to sampling error. Methodology and algorithms are needed to incorporate sampling error components into the existing seasonal adjustment framework.

Current Subprojects:

• Seasonal Adjustment. (McElroy/R&M, Livsey, Pang, Roy)

• Time Series Analysis. (McElroy/R&M, Livsey, Pang, Roy)

• Seasonal Adjustment Software Development and Evaluation. (Livsey, Lytras/ESMD, Tucker McElroy/R&M, Pang, Bell/R&M, Sun/CODS)

Potential Applications:

• Applications encompass the Decennial, Demographic, and Economic areas.

• Avoiding disruption to stakeholders as Economic Directorate moves to Annual Integrated Economic Survey (AIES) design.

• Improved stability to seasonally adjusted data in the presence of outliers or extreme economic shocks such as COVID-19.

• Reduce residual seasonality present in released estimates through improved seasonal diagnostics.

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 2025, staff explored the 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 attempted to assess how much improvement in seasonal adjustment could be achieved using multivariate techniques. Additionally, staff continued developing a weekly seasonal modeling framework using a differencing operator tailored for weekly time series (or more generally, for time series with non-integer periodicities). This framework was applied to further improve seasonal adjustment methods for such data. Staff also continued an evaluation of MSEs in seasonal adjustment and trend estimation using series generated via a component model structure corresponding to airline models with varying parameters and settings.

Staff: Tucker McElroy (R&M), Jim Livsey (x33517), 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 2025, staff worked on developing a methodology that would smoothly transition from univariate to multivariate modeling, while maintaining the familiar characteristics of existing models. This approach would leverage the Final Equation Form of a VAR specification to integrate multivariate techniques while ensuring a user-friendly experience.

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

C. Seasonal Adjustment Software Development and Evaluation

Description: This research is concerned with improvements to software programs performing seasonal adjustment, with the goal of supporting the practice of seasonal adjustment both inside and outside the Census Bureau.

Highlights: During FY 2025, staff documented the development of a Python-based platform for the existing *X-13ARIMA-SEATS* software and RegComponent software.

Staff: Jim Livsey (x33517), Demetra Lytras (ESMD), Tucker McElroy (R&M), Osbert Pang, Bill Bell (R&M), Lijing Sun (CODS), Antoine Devictor (CODS)

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 are 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. In particular, linear mixed effects models are ubiquitous at the Census Bureau through applications of small area estimation. Models can also identify errors in data, e.g., by computing valid tolerance bounds and flagging data outside the bounds for further review.

Research Problems and Potential Applications:

1. Investigate established methods and novel extensions to support design (e.g., factorial designs), analysis, and sample size determination for Census Bureau experiments.
- Sample sizes can be determined to achieve desired power under planned designs and statistical procedures.
- Experimental design can help guide and validate testing procedures proposed for censuses and surveys.

2. Investigate methodology for experimental designs embedded in sample surveys, including large-scale field experiments embedded in ongoing surveys.
- This includes design-based and model-based analysis and variance estimation incorporating the sampling design and the experimental design (van den Brakel, Survey Methodology, 2005).
- Embedded experiments can be used to evaluate the effectiveness of alternative contact strategies, especially for improving response rates.
- Of particular interest to the Census Bureau is where systematic sampling is used both for the sampling design and the experimental design.
- A potential application area is to expand the collection of experimental design procedures utilized with the American Community Survey.

3. Identify and develop statistical models (e.g., loglinear models, mixture models, and mixed-effects models), associated methodologies, and computational tools for problems relevant to the Census Bureau.
- Modeling can help to characterize relationships between variables measured in censuses, sample surveys, and administrative records.
- Modeling can help to study response rates in a census or survey operation and their relationships to associated variables. It can also be used to predict volumes of incoming responses with appropriate measures of uncertainty.
- Models can be used to provide principled measures of statistical variability for constructs like the POP Division's Population Estimates.
- Modeling can enhance information obtained from various sample surveys using auxiliary data sources, such as administrative records.
- Fiducial prediction intervals of random effects can be applied to mixed effects models such as those used in small area estimation.

4. Construct rectangular nonparametric tolerance regions for multivariate data, focusing on multivariate ratio edits.
- This can be applied to multivariate economic data and aid in the editing process by identifying observations that are outlying in one or more attributes and which subsequently should undergo further review.
- The importance of ratio edits and multivariate/multiple edits is noted in the work of Thompson and Sigman (Journal of Official Statistics, 1999) de Waal, Pannekoek and Scholtus (Handbook of Statistical Data Editing and Imputation, 2011), and Ghosh-Dastidar and Schafer (JASA, 2003 and Journal of Official Statistics, 2006).

5. Develop a technique for mis-reporting via the COM-Poisson distribution in order to estimate more accurate count estimates.
- This could be used to assess the amount of misreporting in historical Census datasets to aid in model development to estimate more accurate survey count outcomes.

6. Develop a disclosure policy motivated by the COM-Poisson and related distributions that allows one to protect individual information reported in two-way and multi-way tables.
- This would allow the Census Bureau to release statistical measures associated with a general distributional form while protecting individual privacy.
- This would allow one to estimate the form of multi-way tables of interest while masking the true response data.

Current Subprojects:

• Developing Flexible Distributions and Statistical Modeling for Count Data Containing Dispersion. (Sellers, Morris, Raim)

• Design and Analysis Methods for Experiments. (Mathew, Raim, Sellers)
• Randomization, Re-randomization and Covariate Balance in Treatment-control Comparisons (Ben-David, Mathew)
• Ratio Edits for Multivariate Data Based on Tolerance Rectangles (Mathew)
• Generation of Random Variates for Weighted Distributions. (Raim, Livsey, Irimata)

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 2025, staff revised and resubmitted the 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. Staff are developing a regression model motivated by the Conway-Maxwell-Poisson that allows for excess zeroes and right censoring. Towards the end of FY 2025, limited progress was made due to logistical difficulties for a staff member who is no longer with the Census Bureau. This project is suspended.

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

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

Highlights: During FY 2025, staff carried out an investigation of the different variance estimates available under systematic sampling. Because unbiased estimation of the variance is not possible under a systematic sample, various biased estimators have been suggested in the literature. These include variance estimators based on successive differences and successive difference replicates. Because these estimators are quadratic forms in the sample data, staff investigated the properties of a general nonnegative definite quadratic form to be used as a variance estimator. The investigation was both design-based and model-based. Staff could identify serious issues associated with some of the available variance estimators.

When the population size is a multiple of the sample size, the population values can be considered as a two-way table with the columns representing the different systematic samples. Considering this as two-way classified data, staff noted that *any* reasonable variance estimator is a linear combination of the sum of squares due to the rows and the row x column interaction sum of squares. However, the actual variance is a multiple of the sum of squares due to the columns. This is a clear indication that the performance of a variance estimator depends strongly on the population values. Thus, despite the extensive simulation results reported in the literature, and in the variance estimation book by Wolter, there is no single variance estimate that can be recommended for practical use under systematic sampling. The two-way table based analysis mentioned above can also be used to identify the conditions under which a specific variance estimator is appropriate. However, such conditions are population dependent. Staff is in the process of applying the results to some Census Bureau data and is currently preparing a manuscript that includes the above findings. It is hoped that the results will have applications for variance estimation under some of the embedded experiments carried out at the Census Bureau, where a systematic sample is used both at the sampling stage and at the design stage.

Staff: Thomas Mathew (x35337), Emanuel Ben-David, Andrew Raim

C. 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 2025, there was no significant progress on this project.

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

D. 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, because 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 2025, staff has been investigating the derivation of ratio edit procedures specifically for multivariate data. A standard parametric setup to address the problem is that of multivariate normality. However, even when the individual observation vectors follow a multivariate normal distribution, the ratios do not. Staff pursued two options: (i) use a parametric bootstrap approach under the multivariate normal scenario and (ii) explore a fully non-parametric approach. The latter is attractive in edit applications due to their robustness with respect to parametric distributional assumptions. Classical ellipsoidal edits that are commonly used are built on robust Mahalanobis distances. However, they lack the ratio by ratio limits that editors often prefer, and they can also allow one extreme ratio to be offset by another coordinate. Staff pursued the development of multivariate ratio edits based on rectangular tolerance regions.

From the ratio edit perspective, of particular interest is the rectangular central tolerance region (RCTR). Our extensive numerical experiments showed that the RCTR is a strong default. They produced explicit per ratio bounds that can be coded directly as editing rules. The RCTR kept the Type I error rates under control while lowering missed detections once contamination was present. Of particular interest is the derivation of a "mixed rectangular region"; that is, two-sided tolerance limits are derived for some ratios, and one sided (lower or upper) limits are derived for the others. The latter could be mandated by design or policy. The methods developed have been illustrated using publicly available data from the *Annual Survey of Manufactures* and related economic data. The work was done in collaboration with a researcher (Derek Young) and a graduate student (Daniel Tuyisenge) at the University of Kentucky. A manuscript *Multivariate Ratio Edits Based on Parametric and Nonparametric Tolerance Intervals* based on the above work is currently under preparation.

Staff: Thomas Mathew (x35337)

E. Generation of Random Variates for Weighted Distributions

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 2025, staff completed major revisions of manuscript on vertical weighted strips (VWS) methodology and resubmitted to a journal for peer review.

Staff resumed work on a "vws" R package to support VWS methodology. It uses a C++ interface for fast runtime performance, safer / more formal type handling, and object-orientation. A vignette with detailed API reference and examples was developed with the package. To support programming with the "vws" package, staff submitted an updated version of the "fntl" package to CRAN; this includes density, CDF, quantile, and variate generation functions for univariate distributions which are truncated to intervals.

Staff completed a manuscript applying VWS to a Bayesian small area estimation model. Here, one of the conditionals in the Gibbs sampler is an unfamiliar weighted distribution that can be handled with VWS. To avoid excessive computation from repeated proposal construction within a Gibbs sampler, a "self-tuning" approach is considered where VWS proposals persist during the course of MCMC sampling and knots are altered as needed. Simulations are presented to study tolerances that control the addition and removal of knots to the proposal. An example analysis using public SAIPE data serves as a proof-of-concept of the method. Manuscript was submitted to a journal for peer review.

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

F. A Joint Confidence Region for a Ranking Based on Differences of Ordered Means

Description: This work will contribute to the methodological development of joint confidence regions for a ranking of several populations, based on the novel ideas proposed in Klein, Wright, and Wieczorek (2020), to be referred to as KWW, and in Wright (2024, 2025). In their work, KWW have developed a measure of

uncertainty for an estimated ranking based on confidence intervals for the population means, and by considering the overlap/non-overlap of the confidence intervals. Later, Wright (2024) derived a new joint confidence region for a ranking based on confidence intervals for the population mean differences and by considering the overlap/non-overlap of the intervals with the number zero. It was also noted that this latter approach could be an improvement in terms of the uncertainty. The present work will take up the ideas proposed in Wright (2024, 2025), and will explore improvements in three directions: (i) consider differences among the ordered population means rather than the population means themselves, (ii) construct a rectangular confidence region for the differences among the ordered population means with exact coverage probability, implemented using a parametric bootstrap, and (iii) tighten the joint confidence regions for the ranking by distributing the confidence level unequally among the differences of the ordered population means, with a smaller difference receiving a smaller confidence level. The strategy proposed in (iii) is expected to produce a narrower interval for a smaller difference, reducing the chance of overlapping with the number zero.

Highlights: During FY 2025, staff pursued the ideas proposed in Wright (2024, 2025), and explored possible improvements in three directions: (i) consider differences among the ordered population means rather than the population means themselves, (ii) construct a rectangular confidence region for the differences among the ordered population means with exact coverage probability, implemented using a parametric bootstrap, and (iii) tighten the joint confidence regions for the ranking by distributing the confidence level unequally among the differences of the ordered population means, with a smaller difference receiving a smaller confidence level. The strategy proposed in (iii) is expected to produce a narrower interval for a smaller difference, reducing the chance of overlapping with the number zero. Staff noted that for (i) and (ii), the quantities required are simultaneous lower confidence limits for the differences of the ordered population means, and these can be derived using a parametric bootstrap procedure. Furthermore, the recommendation in (iii) can also be implemented using a parametric bootstrap procedure, after specifying the unequal marginal probabilities. Staff is the process of carrying out simulations in order to assess the performance of the above procedure, and for comparing it with the solutions available in the literature.

Staff: Thomas Mathew (x35337)

Simulation, Data Science, & Visualization

Motivation: Simulation studies that are carefully designed under realistic sample survey or census 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 computationally intensive statistical methods. Modeling also enables one to efficiently use all available information when producing estimates. Such studies can benefit from software for data processing, especially large data sets from nontraditional sources. Data visualizations can help reveal insights. 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 new methods for statistical disclosure control that simultaneously protect confidential data from disclosure while enabling valid inferences to be drawn on relevant population parameters.
- 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).
- Investigate noise infusion and synthetic data for statistical disclosure control.

Current Subprojects:
- Visualizing the United States. (Yau)
- The Ranking Project: Methodology Development and Evaluation. (Wright, Yau, Wieczorek/Colby College, Hall)

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 sample 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 2025, staff designed and developed a framework that allows a user to quickly find details by geography and demographic groups. An information hierarchy and visualization methods focused on data topics allow general users to better relate to the data. Further, staff expanded the framework to allow for dataset additions for additional topics and spans of time. This makes comparisons across geography, categories, and time more straightforward for users now and in future data updates.

Staff: Nathan Yau (CSRM, FLOWINGDATA.COM)

B. The Ranking Project: Methodology Development and Evaluation [See project by same name under the topic Sampling Estimation & Survey Inference.]

C. Inference about a Binomial Proportion under Privacy Protection
Description: In this project, we consider the inferential problem for a Binomial proportion in situations when the exact number of units possessing an attribute under consideration is unavailable due to privacy reasons; however, a synthesized version of this number is available. We consider several perturbations of the original count data and develop appropriate inference methods. We also provide a measure of privacy protection.

Highlights: During FY 2025, staff initially considered the inference problem, addressing it under three types of available information: noise added number and plug-in sampling based and posterior sampling based data. A comparison of the three modes of data source is made based on inferential accuracy and a measure of privacy. Then staff considered the above inference problem and added two more variations of data perturbation methods. Appropriate inference methods and privacy measures have also been developed. This research has been completed and a technical report has been submitted to internal reviewers for processing as a *CSRM Research Report* and ultimate journal publication.

Staff: Bimal Sinha (x34890), Adam Hall, and Nitul Singha (Clarkson 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: Due to funding uncertainty, planning for the *2025 SUMMER AT CENSUS* was canceled.

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

Chen, S., Xu, C., and Cutler, J. (In Press). "Integrating Probability and Non-probability Samples through Deep-Learning-Based mass Imputation," *Survey Methodology*.

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*.

Engmark, J.D. and Opsomer, J.D. (In Press). "Generalized Regression Estimation Under Misspecified Sample Design," *Survey Methodology*.

Hall, A.C. and Kang, J. (2025). "Inference with Pólya-Gamma Augmentation for U.S. Election Law," *Mathematics*, *13*(6), 945. https://doi.org/10.3390/math13060945

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., Livsey, J.A., and Raim, A.M. (2025). "Bayesian Methods to Improve the Accuracy of Differentially Private Measurements of Constrained Parameters," *Journal of Privacy and Confidentiality, 15(2)*. https://doi.org/10.29012/jpc.936

Joyce, P.M. and McElroy, T.S. (2024). "Modeling Survey Time Series Data with Flow-Observed CARMA Processes," *Journal of Official Statistics, 40(4)*, 601-632. https://journals.sagepub.com/doi/full/10.1177/0282423X241286236

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, Vol 12, 5,* 1567-1589, 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),* https://doi.org/10.3390/math12172723

Klein, M. and Sinha, B. (2024). "Multiple Imputation for Parametric Inference Under a Differentially Private Laplace Mechanism," Professor C.R. Rao Memorial Volume of the *International Journal of Statistical Sciences (IJSS), Vol 24(2)*.

Kundinger, B., Reiter, B., and Steorts, R. (2025). "Efficient and Scalable Bipartite Record Matching with Fast Beta Linkage (fabl)," *Bayesian Analysis, 20(3),* 949-972.

Lahiri, S.N., McElroy, T., and Weinberg, D. (2025). "Locally Stationary Spatial Processes," *Sankhya A*, 1-39.

Li, J., Chung, H.C., Okech, D., and Datta, G.S. (2024). "Hierarchical Bayes Small Area Estimation from Aggregated Data using Various Spatial Models," *Statistics and Applications, 22(3)*, 449-469.

Livsey, J. and McElroy, T. (2024). "Applying the Expectation-Maximization Algorithm to Multivariate Signal Extraction," *Journal of Official Statistics 40(4)*, 660-684.

McElroy, T.S., Pang, O.C., and Chen, B. (In Press). "Mitigating Residual Seasonality while Preserving Accounting Relations in Hierarchical Time Series," *Journal of Business and Economic Statistics*.

Shah, N., Basak, J., Sahni, S., Mathur, A., Park, K., Weinberg, D., and Rajasekaran, S. (2025). "Double Metaphone Blocking: An Innovative Blocking Approach to Record Linkage," *2025 International Symposium on Bioinformatics Research and Applications (ISBRA 2025), 15757*, 139-150.

Shah, N., Soliman, A., Basak, J., Sahni, S., Hesse, K., Mathur, A., Park, K., Weinberg, D., White, J., and Rajasekaran, S. (2024). "The Soundex Blocking: A Novel Blocking Approach for Record Linkage," *2024 IEEE International Conference on Big Data (BigData),* 4039-4047.

Slawski, M., West, B.T., Bukke, P., Wang, Z., Diao, G., and Ben-David, E. (2025). "A General Framework for Regression with Mismatched Data Based on Mixture Modeling," *Journal of the Royal Statistical Society Series A, vol 188, Issue 3,*896-919.

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.

Thibaudeau, Y. (2025). "A Review of Modern Multivariate-Derived and Partition-Based Record Linkage Methods (Invited)," *Wiley Interdisciplinary Reviews: Computational Statistics, 17e*, 70015.

West, B.T., Slawski, M., and Ben-David, E. (2025). "Improved Ensemble Predictive Modeling Techniques for Linked Social Media and Survey Data Sets Subject to Mismatch Error," *Methods, Data, Analyses (MDA).* https://doi.org/10.12758/mda.2025.04

Wright, T. (2025). "Optimal Tightening of the KWW Joint Confidence Region for a Ranking," *Statistics and Probability Letters, Vol 217 (6B), 110288,* https://doi.org/10.1016/j.spl.2024.110288

3.2 BOOKS/BOOK CHAPTERS

McElroy, T. and Livsey, J. (2025). "Analysis of Official Time Series with Ecce Signum, an R package for Multivariate Signal Extraction and Forecasting," *Handbook of Statistics – Statistics in Industry and Government, Volume 53*.

Chen, S. and Xu, C. (In Press). "On the Use of Machine Learning Methods for Missing Data Problems (Invited)," *Handbook of Statistics, Statistics in Industry and Government, Vol 53*, Elsevier.

3.3 PROCEEDINGS PAPERS

Joint Statistical Meetings, American Statistical Association, Portland, Oregon, August 3-8, 2024
2024 Proceedings of the American Statistical Association

- Joyce, P.M. and Slud, E., "Statistical Methods Underlying International Migration Estimates at the Census Bureau," https://doi.org/10.5281/zenodo.14020310
- Powers, R., Eltinge, J., Martinez, W., and Morris, D.S. (2024). "Using Linked Micromaps for Evidence-Based Policy," https://doi.org/10.5281/zenodo.14014055
- Weinberg, D. and Mule, T. (2024). "Triple System Estimation of National Population Counts Through Log-linear and Latent Class Modeling," zenodo

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 #2024-07): Adam Hall, "Interpreting and Extending the Maximum Ratio Rest of Unacceptability," October 3, 2024.

RR (Statistics #2024-08): William Bell and Carolina Franco, "Modeling Short Time Series of Cross-Sectional Data for Small Area Estimation," December 23, 2024.

RR (Statistics #2025-01): Adam Hall, "Maximum Norm Ratio Test," April 4, 2025.

RR (Statistics #2025-02): Mary H. Mulry, Thomas Mule, Andrew D. Keller, and Scott M. Konicki, "Using Administrative Records for Enumeration in the 2020 U.S. Census," April 17, 2025.

RR (Statistics #2025-03): Joseph Kang and Adam Hall, "A Machine Learning Approach for Counting Language Minority Groups in the United States," May 8, 2025.

RR (Statistics #2025-04): Daniel Weinberg and Yves Thibaudeau, "Algebraic Dimensional Reduction for the Fellegi-Sunter Model in Record Linkage and General Parameter Input Specification for BigMatch," September 10, 2025.

3.5 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY STUDY SERIES

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

SS (Statistics #2025-01): Kyle Irimata and Tommy Wright, "Statistical Modeling of Reliability for Redistricting Plans," August 27, 2025.

3.6 OTHER REPORTS

Raim, A., Livsey, J., and Irimata, K. (2025). "Rejection Sampling with Vertical Weighted Strips," https://arXiv.org/abs/2401.09696.

Raim, A., Irimata, K., and Livsey, J. (2025). "Self-Tuned Rejection Sampling within Gibbs and a Case Study in Small Area Estimation," https:doi.org//10.48550/arXiv.2509.17155.

Raim, A. (2025). "allocation: Exact Optimal Allocation Algorithm for Stratified Sampling," https://cran.r-project.org/package=allocation.

4. TALKS AND PRESENTATIONS

50th Anniversary of the Department of Statistics Conference, The Ohio State University, Columbus, Ohio, October 7, 2024.
- Tommy Wright (Keynote Address), "(Remembrances and) Data at Foundation of America's Democracy."

International Conference on Advances in Interdisciplinary Statistics and Combinatorics (AISC), The University of North Carolina at Greensboro, Greensboro, North Carolina, October 11, 2024.
- Kimberly Sellers (Invited Speaker), "Multivariate Approaches Towards Modeling Count Data."

2025 Joint Mathematics Meetings, Seattle, Washington, January 10, 2025.
- Tommy Wright (Invited talk in AMS Special Session on the Mathematics of Decisions, Elections, and Games), "Equal Proportions: 'As Near As May Be.'"

International Conference on Advanced Data Analytics and Statistics, Vimala College, Kerala, India, January 23-25, 2025.
- Thomas Mathew (Invited talk), "Reference Intervals and Regions in Laboratory Medicine."

Space Time Analysis Bayesian Research Group Seminar, University of Washington, Seattle, Washington, April 29, 2025.
- Gauri Datta (Invited Virtual Talk), "Credible Regions for Ranks of Entities

Online Seminar, Department of Applied Mathematics & Statistics, State University of New York at Stony Brook, Stony Brook, New York, U.S., May 30, 2025.
- Joseph Kang, "On Machine Learning Models for Incomplete Survey Data."

The Government Advances in Statistical Programming (GASP) 2025 Virtual Conference, Washington, D.C., June 25, 2025.
- Joseph Kang (and Adam Hall), "A Machine Learning Approach for Counting Language Minority Groups in the United States."

2025 Tenth International Webinar on "Recent Trends in Statistical Theory and Applications," Department of Statistics, University of Kerala, Kerala, India, June 29-July 2, 2025.
- Gauri Datta (Plenary Session), "Credible Distributions of Overall Ranking of Entities."

2025 Small Area Estimation Conference, Turin, Italy, July 7-10, 2025.
- Gauri Datta (Invited Talk), "Credible regions for Ranks of Entities."

2025 Joint Statistical Meetings, Nashville, Tennessee, August 2-7, 2025.
- Gauri Datta, "Credible Regions for Ranks of Entities."
- Andrew Raim (Invited talk), "Design and Analysis of an Experiment for Nonresponse Follow-up in the 2020 Decennial Census."
- Darcy Morris (Invited talk), "Model-Based Weighting for Nonresponse in the American Community Survey: Evaluation and Visualization."

Online Seminar, Department of Mathematics and Statistics, Indian Institute of Technology (IIT), Kanpur, India, August 22, 2025.
- Bimal Sinha, "Confidence Ellipsoids of a Multivariate Normal Mean Vector Based on Noise-Perturbed & Synthetic Data with Applications."

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

Emily Peterson, Emory University, "A Bayesian Hierarchical Small Area Population Model Accounting for Heterogenous Data Source Specific Methodologies from ACS, PEP, and Census Data," October 16, 2024.

Mark Meyer, Georgetown University, "Bayesian Wavelet-Packet Historical Functional Linear Models," October 30, 2024.

Yves Thibaudeau, U.S. Census Bureau, "A Review of Modern Multinomial-Derived and Partition-Based Record-Linkage Methods," December 3, 2024.

Tommy Wright, U.S. Census Bureau, "Uncertainty Reduction for an Estimated Ranking Using Differences," June 25, 2025.

Gauri Datta, University of Georgia/U.S. Census Bureau, "Credible Distributions of Overall Ranking of Entities," June 26, 2025.

Anindya Roy, University of Maryland, Baltimore County/U.S. Census Bureau, "Maximum Entropy Implementation of Differential Privacy Under Aggregation Constraints," July 15, 2025.

Joseph Kang & Adam Hall, U.S. Census Bureau, "A Machine Learning Approach for Counting Language Minority Groups in the United States," July 21, 2025.

Sauvar Guha, Bihar Agricultural University, "Small Area Estimation Under a Spatially Correlated Multivariate Area-Level Model," August 5, 2025.

Sixia Chen, ASA/NSF/Census Research Fellow, The University of Oklahoma Health Sciences, "Introduction to Probability and Nonprobability Sampling" (Lecture 1 of 5 Lectures on Analytic Tools for Non-probability Samples), August 19, 2025.

Sixia Chen, ASA/NSF/Census Research Fellow, The University of Oklahoma Health Sciences, "Calibration Methods" (Lecture 2 of 5 Lectures on Analytic Tools for Non-probability Samples), August 26, 2025.

Sixia Chen, ASA/NSF/Census Research Fellow, The University of Oklahoma Health Sciences, "Propensity Score Methods" (Lecture 3 of 5 Lectures on Analytic Tools for Non-probability Samples), September 9, 2025.

Sixia Chen, ASA/NSF/Census Research Fellow, The University of Oklahoma Health Sciences, "Mass Imputation Methods" (Lecture 4 of 5 Lectures on Analytic Tools for Non-probability Samples), September 16, 2025.

Sixia Chen, ASA/NSF/Census Research Fellow, The University of Oklahoma Health Sciences, "Doubly Robust and Multiply Robust Methods" (Lecture 5 of 5 Lectures on Analytic Tools for Non-probability Samples), September 23, 2025.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

2024 (Inaugural) Alumni Award, Department of Statistics, The Ohio State University

- Tommy Wright

6.2 SIGNIFICANT SERVICE TO PROFESSION

Serge Aleshin-Guendel

- Refereed papers for *Annals of Applied Statistics, Journal of Survey Statistics and Methodology, Journal of the Royal Statistical Society Series A, Journal of the Royal Statistical Society Series B, and Statistical Methods & Application*
- Member, PhD in Statistical Science Committee, Duke University

Emanuel Ben-David

- Member, Program Committee, 18th (2025) International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction & Behavior Representation in Modeling & Simulation
- Refereed papers for the *Annals of Applied Statistics* and the *Statistics and Computing Journal*

Gauri Datta

- Associate Editor, *Sankhya*
- Associate Editor (as outgoing member), *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 Survey Statistics and Methodology (2), Biometrika, Metron, Journal of the American Statistical Association*

Kyle Irimata

- Refereed papers for *Journal of Survey Statistics and Methodology*, *Journal of the Royal Statistical Society Series A,* and *BMC Medical Research Methodology*

Ryan Janicki

- Refereed papers for *The R Journal*, *Communications in Statistics, Canadian Journal of Forest Research, JMIR Public Health and Surveillance, Journal of Statistical Computation and Simulation, Journal of Survey Statistics and Methodology*

Patrick Joyce

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

Joseph Kang

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

James Livsey

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

Jerry Maples

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

Thomas Mathew

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

- Chair, W.J. Youden Award in Interlaboratory Testing Committee, American Statistical Association
- Guest Editor, Special Issue of the Journal of *Statistics and Applications* in honor of CR Rao
- Guest Editor, Special Issue of the *International Journal of Statistical Sciences* in honor of CR Rao

Darcy Morris
- Associate Editor, *Communications in Statistics*
- *Newsletter* Editor, Survey Research Methods Section, American Statistical Association
- Program Chair, Government Statistics Section, American Statistical Association
- Member, External Nominations and Awards Committee, American Statistical Association
- Refereed a paper for *Journal of the Royal Statistical Society – Series A*

Mary Mulry
- Associate Editor, *Journal of Official Statistics*

Tapan Nayak
- Associate Editor, *Journal of Statistical Theory and Practice*
- Guest Editor, Special Issue of *Statistics and Applications* in Memory of Professor C.R. Rao

Osbert Pang
- Refereed a paper for *Data Science in Science*

Andrew Raim
- Refereed papers for *Communications in Statistics – Theory and Methods*, *Statistics in Medicine*, and *Statistics and Public Policy*
- Chair, Session, 2025 Joint Statistical Meetings, American Statistical Association

Kimberly Sellers
- Associate Editor, *The American Statistician*
- Guest Editor, *Applied Stochastic Models in Business and Industry*
- Chairperson, External Nominations and Awards Committee, American Statistical Association
- Member, Committee on Applied and Theoretical Statistics, National Academies of Sciences, Engineering, & Medicine
- Member, Deming Lectureship Committee, American Statistical Association

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*

Rebecca Steorts
- Associate Editor, Bayesian Analysis
- Associate Editor, *Journal of Survey Statistics and Methodology*
- Associate Editor, *Journal of the American Statistical Association, Applications and Case Studies*

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

6.3 PERSONNEL NOTES

- Tapan Nayak ended his Schedule A Appointment December 10, 2024 following 16 years of federal service.
- Eric Slud retired January 31, 2025 following 13 years (plus 10 years with a Schedule A Appointment) of federal

service.

- Paul Parker's Schedule A Appointment ended March 8, 2025.
- Tom Petkunas retired March 8, 2025 following 40+ years of federal service.
- Chad Russell retired March 8, 2025 following 34 years of federal service.
- Michael Hawkins retired May 2, 2025 following 44+ years of federal service.
- Kimberly Sellers ended her Schedule A Appointment May 22, 2025 following 1 year as an ASA/NSF/Census Research Fellow plus nearly 10 years of federal service.
- Jim Livsey ended his Census Bureau appointment July 2, 2025 following 12^{-} years of federal service.
- Mary Mulry retired August 1, 2025 following 40 years of federal service.

APPENDIX A **Center for Statistical Research and Methodology FY 2025**
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		
5350M04	Person Characteristic Frame		
5450M20	In-Office Enumeration		
5450M21	Response Quality Assurance Follow-Up		
5450M23	Response Processing Planning & Support		
5550M01	Data Product Creation and Dissemination		
5550M02	Redistricting Data Program		
5650M01	Evaluations, Experiments, & Research		
5650M02	Post Enumeration Survey Design & Estimation		
	1. Study on Children 10-14 Years Old Counted in the 2020 Census but Uncounted in the 2010 Census	Michael Ikeda	Eric Jensen (Jordan Misra)
	2. Statistical Modeling to Augment 2020 Disclosure Avoidance System	Kyle Irimata	Michael Walsh
	3. 2030 Characteristic Imputation Modeling Research for Nested Data with Structural Zeros	Emanuel Ben-David	Tom Mule
	4. Mobile Questionnaire Assistance: Analysis and Simulation	Andrew Raim	Megan Parker
	5. Continuous Count Study	Michael Ikeda	Tom Mule
	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. Capture-Recapture Coverage Measurement using Administrative Records (Continuous Count Study)	Dan Weinberg	Tom Mule
	8. Cohort Component Birth Modeling (Continuous Count Study)	Emanuel Ben-David	Tom Mule
	9. Cohort Component Domestic Migration Modeling (Continuous Count Study)	Ryan Janicki	Tom Mule
	10. Record Linkage Support for Decennial Census	Ned Porter	Aaron Gilary
	11. Cohort Component International Migration Modeling (Continuous Count Study)	Patrick Joyce	Tom Mule
	12. Coverage Measurement Research	Jerry Maples	Tim Kennel
	13. Agreements for Advancing Record Linkage	Rebecca Steorts	Jennifer Hutnick (Krista Park)
	14. Cohort Component Death Modeling (Continuous Count Study)	Patrick Joyce	Tom Mule
6385M70	American Community Survey (ACS)		
	15. Assessing and Enhancing the ACS Experimental Weighting Approach Implemented in 2020 Data Products	Darcy Morris	Eddie Castro
	16. Voting Rights ACT (VRA) Section 203 Research Towards 2026 Determinations (also Decennial Project 6550J06)	Joseph Kang	James Whitehorne
	DEMOGRAPHIC		
TBA	Demographic Statistical Methods Division (DSMD) Special Projects		
	17. Nonresponse Adjustments for High Frequency Low Response Surveys	Darcy Morris	John Chestnut
0906/1444X00	Demographic Surveys Division (DSD) Special Projects		
	18. Using Machine Learning for Improving Nonresponse Adjustment in the Current Population Survey	Emanuel Ben-David	Tim Trudell
	19. Data Integration	Ned Porter	Chris Boniface
7165025	Social, Economic, and Housing Statistics Division Small Area Estimation Projects		
	20. Research for Small Area Income and Poverty Estimates (SAIPE)	Jerry Maples	Wes Basel
	21. Assessing Constant Parameters across Areas in the SAIPE Models	Jerry Maples	Wes Basel

1183X01 1183X90	**ECONOMIC** General Economic Statistical Support General Economic Statistical Program Management	
	22. Small Area Estimation for the Annual Integrated Economic Survey	Serge Aleshin-Guendel .....Stephen Kaputa
	23. Seasonal Adjustment Support.....	Osbert Pang..... Kathleen McDonald-Johnson
	24. Seasonal Adjustment Software Development and Evaluation	Osbert Pang..... Kathleen McDonald-Johnson
	25. Research on Seasonal Time Series - Modeling & Adjustment Issues.....	Osbert Pang..... Kathleen McDonald-Johnson
	26. Supporting Documentation & Software for Seasonal Adjustment	Osbert Pang..... Kathleen McDonald-Johnson
0331000	**PROGRAM DIVISION OVERHEAD** *27. Research Computing*.....	Chad Russell ..... Jaya Damineni
7225157	**NATIONAL CANCER INSTITUTE** *28. Modeling Tobacco Use Outcomes with Data from Tobacco Use Supplement – Current Population Survey*.....	Isaac Dompreh ..... Benmei Liu

APPENDIX B

FY 2025 PROJECT PERFORMANCE MEASUREMENT QUESTIONNAIRE

CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY

Dear

As a sponsor for the FY 2025 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 2025 ***(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 2025 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 2025 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

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
Joseph Kang (Acting)
- Ned Porter

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

MATHEMATICAL STATISTICS AREA
Jerry Maples (Acting)

Sampling Estimation & Survey Inference Research Group
Tommy Wright (Acting)
- Sixia Chen (ASA/NSF/Census Research Fellow)
- Mike Ikeda
- Patrick Joyce

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.)

Time Series & Seasonal Adjustment Research Group
Jerry Maples (Acting)
- Osbert Pang
- Anindya Roy (UMBC)

Experimentation, Prediction, & Modeling Research Group
Tommy Wright (Acting)
- Thomas Mathew (UMBC)
- Andrew Raim

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