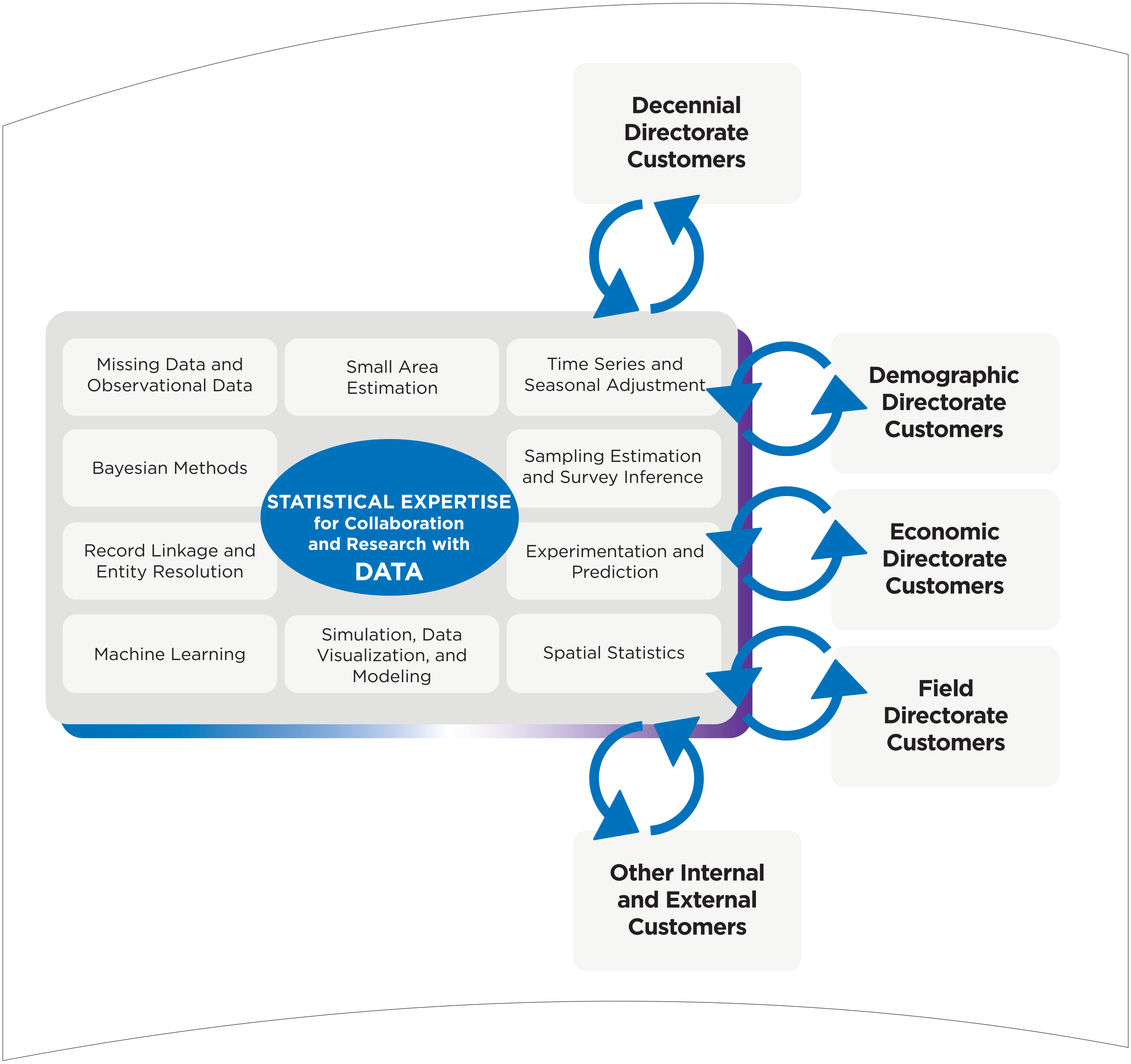
Annual Report of the Center for Statistical Research and Methodology

Research and Methodology Directorate

Fiscal Year 2020



Decennial Directorate Customers

Missing Data and Observational Data

Small Area Estimation

Time Series and Seasonal Adjustment

Bayesian Methods

STATISTICAL EXPERTISE for Collaboration and Research with DATA

Sampling Estimation and Survey Inference

Record Linkage and Entity Resolution

Experimentation and Prediction

Machine Learning

Simulation, Data Visualization, and Modeling

Spatial Statistics

Demographic Directorate Customers

Economic Directorate Customers

Field Directorate Customers

Other Internal and External Customers

United States® Census Bureau

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

Since August 1, 1933—

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

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

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

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

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

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

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

We help the Census Bureau improve its processes and products. For fiscal year 2020, this report is an accounting of our work and our results.

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

Highlights of What We Did...

As a technical resource for the Census Bureau, each researcher in our center is asked to do three things: *collaboration/consulting*, *research*, and *professional activities and development*. We serve as members on teams for a variety of Census Bureau projects and/or subprojects.

Highlights of a selected sampling of the many activities and results in which the Center for Statistical Research and Methodology staff members made contributions during FY 2020 follow, and more details are provided within subsequent pages of this report:

- *Co-published improved statistical models for prediction of self-response rates to the 2020 Census at the tract level combining data from three sources:* (1) self-response results from the 2015 Census Test, (2) Low Response Score from the Planning DataBase, and (3) Tapestry, a third-party population and geographic segmentation.

- *Worked with Decennial Statistical Studies Division to explore the capability of machine learning models to produce a multi-class re-classifier to assign Census mindsets to respondents in new samples selected for an evaluation of the 2020 Census Communication Campaign.*

- *Completed a paper on continuation-ratio logit modeling for sample size determination and analysis of experiments involving sequences of success/failure trials; such models support the study of nonresponse probabilities under multiple enumeration attempts to each household in the 2020 Census;* methodology was used in collaborative research on a bilingual training experiment with the Center for Behavioral Science Methods.

- *Started planning and statistical modeling work for the Voting Rights Act, Section 203 model evaluation and enhancements towards the 2021 determination of which jurisdictions should provide voting materials in languages other than English*: (1) explored several sources of auxiliary information; (2) studied the use of temporal small area estimation models that borrow strength from previous vintages of the American Community Survey.

- *Developed a Bayesian unit-level small area model, which uses a survey-weighted pseudolikelihood, for use with data obtained from an informative survey to predict the number of persons in different categories with health insurance at the county level using Small Area Health Insurance Estimates (SAHIE) Program data.*

- *Worked with Economic Indicators Division and Economic Statistical Methods Division on imputation models using establishment-level data from a commercial retail data aggregator (NPD) to produce an experimental data product of retail sales volume by 3-digit NAICS codes and state on a monthly basis.*

- *Worked with Economic Statistical Methods Division to explore the use of new seasonal adjustment results with the M3 sample survey: comparing X-11 adjustments (current production method) with SEATS adjustment and multivariate signal extraction with latent component models;* important focus on documenting exact SEATS admissible decompositions compared with ARIMA model inputs.

- *Released a study documenting that variability in the TopDown Algorithm of the Disclosure Avoidance System increases as levels of geography and population decrease.*

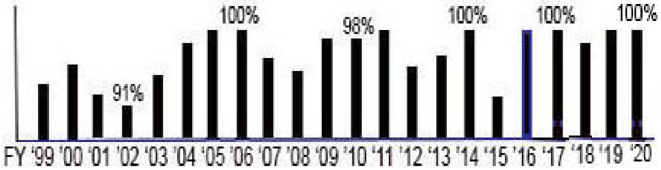
- *GENERAL STATISTICAL RESEARCH: (1) Conducted simulation studies and observed lower bias for a bootstrap approach as compared to traditional estimators of MSE for small area models; (2) Developed novel optimization-based weighting adjustment methods based on partially missing data, along with diagnostics based on cross-classified post-stratification variables; (3) Developed and published a general exact optimal sample allocation algorithm with bounded cost and bounded stratum sample sizes; (4) Wrote EM algorithm for estimating the weights of the Fellegi-Sunter record-linkage model for use with "BigMatch" and the R "RecordLinkage" package, published a joint Bayesian method to utilize blocking and entity resolution, and began to develop a record-linkage infrastructure for the Census Bureau; (5) Developed and implemented new algorithms for ragged edge missing value imputation, and ad hoc filtering of multivariate time series; (6) Completed paper on spatio-temporal change of support modeling in R; and (7) Published a method for controlling identification risks while well preserving the weighted estimates from a sample survey.*

How Did We[1] Do...

For the 22nd year, we received feedback from our sponsors. Near the end of fiscal year 2020, our efforts on 29 of our program (Decennial, Demographic, Economic, Administration, External) sponsored projects/subprojects with substantial activity and progress and sponsor feedback (Appendix A) were measured by use of a Project Performance Measurement Questionnaire (Appendix B). Responses to all 29 questionnaires were obtained with the following results (The graph associated with each measure shows the performance measure over the last 22 fiscal years):

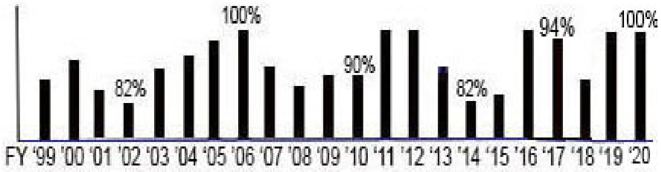
Measure 1. Overall, Work Met Expectations

Percent of FY2020 Program Sponsored Projects/Subprojects where sponsors reported that overall work met their expectations (agree or strongly agree) (27 out of 27 responses) ……..... 100%



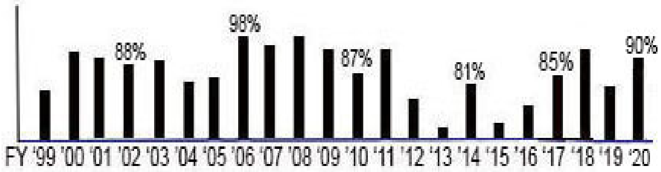
Measure 2. Established Major Deadlines Met

Percent of FY2020 Program Sponsored Projects/Subprojects where sponsors reported that all established major deadlines were met (17 out of 17 responses) ……......…...…...….. 100%



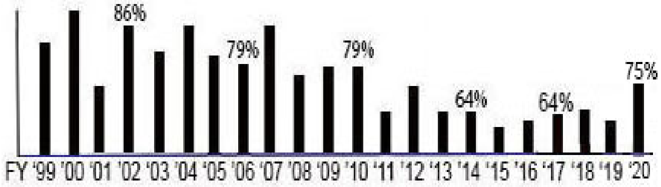
Measure 3a. At Least One Improved Method, Developed Technique, Solution, or New Insight

Percent of FY2020 Program Sponsored Projects/Subprojects reporting at least one improved method, developed technique, solution, or new insight (26 out of 29 responses) ………… 90%



Measure 3b. Plans for Implementation

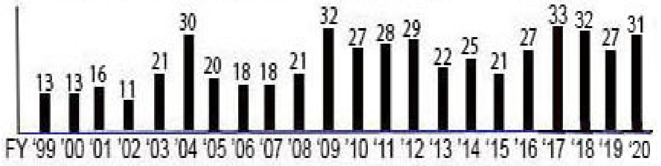
Of these FY2020 Program Sponsored Projects/Subprojects reporting at least one improved method, technique developed, solution, or new insight, the percent with plans for implementation (21 out of 28 responses) …………..…….. 75%



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

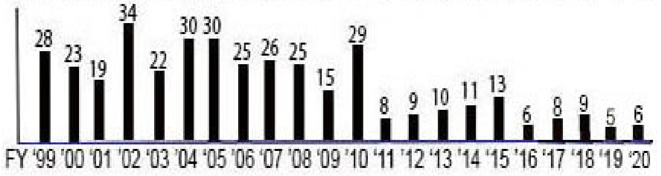
Measure 4. Journal Articles, Publications

Number of peer reviewed journal publications documenting research that appeared (22) or were accepted (9) in FY2020 ……………………………………………………………... 31



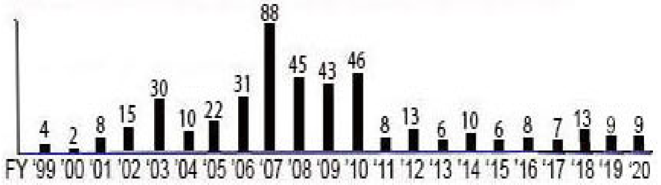
Measure 5. Proceedings, Publications

Number of proceedings publications documenting research that appeared in FY2020 ………………………………..………. 6



Measure 6. Center Research Reports/Studies, Publications

Number of center research reports/studies publications documenting research that appeared in FY2020 ………...…… 9



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

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

TABLE OF CONTENTS

1. COLLABORATION

1.1 ADDRESS CANVASSING IN OFFICE (DECENNIAL Project 6350F01)

1.2 ADVERTISING CAMPAIGN (Decennial Project 6450F20)

1.3 DATA CODING/EDITING/IMPUTATION (Decennial Project 6550F01)

1.4 REDISTRICTING DATA PROGRAM (Decennial Project 6550F06)

1.5 CCM PLANNING & PROJECT MANAGEMENT (Decennial Project 6650F01)

1.6 2020 EVALUATIONS – PLANNING & PROJECT MANAGEMENT (Decennial Project 6650F20)

1.7 2030 PLANNING & PROJECT MANAGEMENT (Decennial Project 6650F25)

1.8 ADMINISTRATIVE RECORDS DATA (Decennial Project 6750F01)

A. 2020 Census Communications Campaign Statistical Analyses - I

Description: Both the 2000 and 2010 U.S. Censuses included a social marketing communications campaign that aided in maintaining the mail response rate in an environment when response to surveys was declining. As the 2020 U.S. Census is underway, the preparations include tests of new methodologies for enumeration that have the potential to reduce cost and improve quality. In parallel, the research included formulating methods for the 2020 Census communications campaign that would aid the effectiveness of the enumeration operations. A team has been set up to conduct the research. For example, the 2015 Census Test in Savannah, GA included tests of Internet and mail response modes and of online delivery of social marketing communications focused on persuading the public to respond by Internet and mail. Analyses of the 2015 Census Test results and other data supported the preparations for the 2020 Census communications campaign.

Highlights: During FY2020, staff continued to collaborate with staff in the Research and Methodology Directorate to explore the use of a lifestyle segmentation of the U.S. population to gain insight about variation in self-response in censuses and sample surveys and determine whether the segments could be useful in the 2020 Census Communications Campaign. The use of the Tapestry segmentation is innovative in that it is designed for commercial marketing and not commonly used in sample survey research or census taking. The initial analyses used merged data from three sources: (1) self-response results from the 2015 Census Test in Savannah, GA, (2) the Low Response Score (LRS) found on the Census Bureau's Planning Database, and (3) Tapestry, a third-party population and geographic segmentation to create a dataset suitable for studying relationships between census response, the LRS, and lifestyle segments. The models show that adding lifestyle segments to the LRS improves the prediction of self-response rates at the tract level. A manuscript that describes the results have been accepted for publication in the *Journal of Survey Statistics and Methodology.*

Staff: Mary Mulry (682-305-8809)

A. 2020 Census Communications Campaign Statistical Analyses - II

Description: The U.S. Census Bureau has fielded the 2020 Census Communications Campaign to encourage participation in the 2020 Census. Similar campaigns aided in maintaining high self-response rates for the 2000 and 2010 Censuses. To prepare, the U.S. Census Bureau fielded the 2020 Census Barriers, Attitudes and Motivators Survey (CBAMS) to collect data on attitudes and knowledge about the U.S. Census. Data from over 17,000 respondents was used to classify individuals into one of six mindsets that could be used in developing messages to persuade individuals to respond.

Highlights: During FY2020, staff worked with staff in the Decennial Statistical Studies Division to conduct research to identify the type of multi-class classification model that produced the best re-classifier for assigning 1 of 6 Census mindsets to respondents in new samples selected for an evaluation of the 2020 Census Communications Campaign. The evaluation, called the Mindset Shift Study, examines whether Census mindsets changed during the Campaign. The CBAMS data were transformed into eight factors using Principal Components Analysis to develop the mindsets. However, the clustering method used in forming the Census mindsets was not suitable for classifying respondents who answered the same questionnaire in new samples. The team used the CBAMS PUMS data and followed the same procedure to create the same 8 factors so they could replicate the formation of the mindsets as close as possible. These 8 factors were used as independent variables when fitting the models so the process did not involve variable selection. The plan included fitting the selected type of model using the CBAMS data and using it to assign mindsets to the new samples for analyses.

The team considered five types of models: linear discriminant analysis, quadratic discriminant analysis, multinomial logistic regression, random forest, and support vector machine (SVM). Both random forest and SVM required tuning to find the optimal settings for a pair of parameters. The SVM had the highest value of each of the three evaluation criteria discussed below.

The research also examined whether the multi-class AUC, an evaluation method for multi-class classification models developed by Hand and Till (2001), was helpful in identifying the new classifier for assigning mindsets. The multi-class AUC is a generalization of the area under the curve (AUC) for receiver operating curves (ROC) for binary classification models. The multi-class AUC was found to be a helpful supporting measure in evaluating the 5 types of models, particularly when used in combination with other evaluation criteria. Alongside the overall correct classification rate and the range of the correct classification rates for the individual mindsets, the multi-class AUC aided in weighing the strengths and weaknesses for each model.

The team's research is innovative in that it explored the capability of machine learning models to produce a multi-class re-classifier to assign Census mindsets to respondents in new samples selected for an evaluation of the 2020 Census Communications Campaign. The results provided an improvement in methodology over the approach used in a similar study that evaluated the 2010 Census Integrated Communications Program (ICP). The results of the study are in "Multi-class Classification Model for Census Mindsets" that was submitted to the *Proceedings of the 2020 Joint Statistical Meetings.*

Staff: Mary Mulry (682-305-8809), Darcy Steeg Morris, Isaac Dompreh

B. Supplementing and Supporting Non-response 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 FY2020, staff ran several implementations of an outlier detection methodology on IDs in two Master Address File (MAF) extracts. Staff began by rerunning an implementation developed by Decennial Statistical Studies Division staff, then updated the implementation for a more recent MAF extract, then further modified the programs in multiple stages. Key modifications include using a more complete (in terms of variables present) version of the updated MAF extract, using an updated version of the Delivery Sequence File (DSF) status flag, adding action codes from in-field address canvassing, LUCA action codes, an indicator for the presence of administrative records (AR) persons, adding supplemental NRFU IDs to the modeling, adding an indicator for supplemental NRFU ID, and a categorical variable based on information from undeliverable as addressed (UAA) nixie codes to the modeling variables. A file of the top 100 tracts (500+ units in self response TEA) by ratio object score was sent to Geography Division. Staff continued to update the outlier detection results using updated UAA information and AR presence information. In general, high ratio object scores (the score from the outlier detection methodology) tend to be associated with types of enumeration area (TEA) other than self-response or with indications of questionable ID quality. Staff also found that tracts with larger shortfalls in the tax year 2019 IRS 1040 response rate (compared to tax year 2018) tended to have lower tract medians and tract third quartiles of the ratio object score. Staff compared the results from the outlier detection methodology to the results from the 2020 Decennial production AR modeling methodology. High ratio object scores are associated with assigned AR delete status and (to a lesser extent) with assigned AR vacant status. This effect is more pronounced for IDs in the self-response TEA (TEA 1) than for IDs in the update/leave TEA (TEA 6). These effects were generally consistent across different iterations of the AR modeling runs. Later comparisons also looked at former AR vacant/delete overlap units converted to AR vacants or AR deletes and at AR closeout cases (AR assigned status with relaxed conditions for use during NRFU closeout). In general, former overlap cases converted to AR deletes looked somewhat similar to regular AR deletes in their ratio object score distribution, although there were clear differences in the distribution of some outlier detection modeling variables. Former overlap cases converted to AR vacants were in between regular AR deletes and regular AR vacants in their ratio object score distribution in TEA 1 and somewhat similar to regular AR deletes in TEA 6. Closeout cases in TEA 1 looked fairly similar in ratio object score distribution to non-closeout cases with the corresponding AR status assignment. Closeout vacants, deletes, and vacant/delete overlap cases in TEA 6 looked similar to each other in the ratio object score distribution and in-between the ratio object score distributions for non-closeout AR vacants and AR deletes in TEA 6. Staff presented overviews of the outlier detection results to the AR Modeling sub-team and the AR Usage team.

Staff also examined the 2008 and 2009 Puerto Rico tax files and compared them to a 2010 AR modeling file to explore the possible usefulness of the Puerto Rico tax file for AR modeling in Puerto Rico. The Puerto Rico tax files had a very low proportion of nonblank MAF IDs, and the household size from the tax file did not appear to be strongly related to the Census household size. Some additional persons could be obtained by using the PIK to match to existing AR persons and then borrowing the MAF ID from the existing AR persons. However, this would also be of limited use because of the low proportion of 2020 Puerto Rico MAF IDs that have AR records present. Staff concluded that it did not appear to be worthwhile to pursue a new data agreement to acquire current Puerto Rico tax file information for the purposes of AR modeling. If a current Puerto Rico tax file were acquired for some other reason, it might be of some use for research into both modeling household status and adding additional persons to households. Staff provided a document outlining the Puerto Rico tax file work to the AR Usage team.

Finally, staff examined processed files of off-campus residents obtained from various colleges, focusing on indicators relevant to potential duplication. The files could have both local (near-campus) and alternate (possibly parental) addresses, although only a relatively small proportion of records on the files included two addresses that could be matched to two different MAFIDs. There was some (although not a large proportion) indication of duplication (based on persons from the files matching to 2020 Census self-response persons from two different MAFIDs) and some (although not a large proportion) of this duplication was between the local and alternate addresses. Staff presented results to the AR Usage team, with updates as files from additional colleges were processed.

Staff: Michael Ikeda (x31756)

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

Highlights: During FY2020, methodology was developed to produce realistic multiway arrays for simulated population counts. These arrays will act as ground truth and represent a logical way to induce realistic correlation into our simulation study. The new methodology allows users to control the amount of cross sectional dependence present in a synthetic population table.

Staff: James Livsey (x33517), Eric Slud

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

Highlights: During FY2020, staff worked with Decennial Statistical Studies Division and Center for Economic Studies colleagues on: contingency plan exploratory analysis to assess model impacts of NRFU delays and the extension of the IRS filing deadline; adaptation of models for identifying and enumerating occupied housing units on American Indian reservations; and adaptation of models for identifying and enumerating occupied housing units for off-campus college/university housing units.

Staff: Darcy Steeg Morris (x33989), Yves Thibaudeau

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

Highlights: During FY2020, staff used the proposed sample size determination method to justify that a preselected set of area census offices (ACOs) would be sufficient for the bilingual training experiment. The recommendation included a maximum number of contact attempts to consider in the analysis, to avoid a situation where data are too sparse to rely on large sample results. Staff prepared a research report based on the methodology, restricted to the case of fixed effects, and featuring the bilingual training experiment as a motivating illustration. The manuscript was also submitted to a journal for peer review.

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

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

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

Highlights: During FY2020, a further meeting was held with decennial administrative-records experts to suggest the relevance of Numident administrative records to the machine-learning approach to detecting and predicting MAF changes among existing housing stock. Staff took some steps to make these data available. In further correspondence, it was decided that the machine-learning predictive approach to MAF changes will be written up as a separate part of the Decennial Evaluation project on MAF updating.

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

G. Record-Linkage Support for the Decennial Census

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

Highlights: During FY2020, staff continued to provide record-linkage infrastructure for the 2020 Census as well as laying the ground for long-term efforts in record linkage at the Census Bureau. Staff began exploring several open-source and commercial (Senzing) softwares. The goal is to construct an integrated record-linkage infrastructure at the Census Bureau for use in many sample surveys and Censuses. Staff has begun a long-term effort at the Census Bureau to develop a record-linkage infrastructure capable of supporting operations traditionally supported by field payloads. The objective is to have an integrated system in place and functional several years before the 2030 Census. In that context, staff explored existing software in R (RecordLinkage, fastLink), C (BigMatch) and Python (FEBRL, MAMBA) and gave multiple presentations and recommendations written in a report on how such software applications could serve the long-term needs of the Census Bureau. The focus is on unduplicating and counting persons and households across multiple administrative lists and commercial files. In the context of the 2020 Census, staff linked Census person files including upward of 400 Million records to commercial files to identify and get estimates of the number of persons without "PIKs" in the U.S. Staff assisted in matching off-campus addresses for college students to the MAF. This is especially important this year (2020) because many college students left their universities in March, due to the ongoing pandemic. Staff investigated and adapted the "MAMBA" business matching methodology from the Economic Directorate to linking files of persons through the matching of surnames, given names, as well as of addresses.

Staff: Yves Thibaudeau (x31706), Emanuel Ben-David, Daniel Weinberg, Rebecca Steorts, David Brown (CES)

H. Coverage Measurement Research

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

Highlights: During FY2020, staff helped with the final reviews of specification documents for the 2020 Post Enumeration Survey (PES). Staff joined the newly created Post Enumeration Survey (PES) Quality Team in response to the Covid-19 outbreak to determine the potential impacts to the PES operations The team produced two documents: *DSSD 2020 POST-ENUMERATION SURVEY MEMORANDUM SERIES* "#2020-A-03 Impacts of COVID-19 on the Post-Enumeration Survey" and "#2020-A-04 PES Quality Team Response to COVID-19 Changes" which discuss the additional risks that the COVID-19 pandemic brings to the PES procedures and suggestions for how the PES can adapt and change to deal with these issues while trying to complete its mission.

Staff: Jerry Maples (x32873), Ryan Janicki, Eric Slud

I. Development of Block Tracking Database

Description: The Targeted Address Canvassing (TRMAC) project supports Reengineered Address Canvassing for the 2020 Census. The primary goal of the TRMAC project is to identify geographic areas to be managed in the office (i.e., in-office canvassing) and geographic areas to be canvassed in the field. The focus of the effort is on decreasing in-field and assuring the Master Address File (MAF) is current, complete, and accurate. The Block Assessment, Research, and Classification Application (BARCA) is an interactive review tool which will allow analysts to assess tabulation blocks—and later Basic Collection Units (BCUs)—by comparing housing units in 2010 imagery and current imagery, along with TIGER reference layers and MAF data.

Highlights: During FY2020 and after reaching the In-Office Address Canvassing (IOAC) Interactive Review goal of completing 100% of all 11 million Census tabulation blocks in the U.S., the Census Bureau only needed to field canvass 35% of all addresses and approximately 20% of all Basic Collection Units (BCU) in the U.S. This project has been completed.

Staff: Tom Petkunas (x33216)

J. A Deterministic Retabulation of Pennsylvania Congressional District Profiles from the 115th Congress to the 116th Congress

Description: During the 115th Congress and before the start of the 116th Congress, congressional districts in Colorado and Pennsylvania were redrawn. Changes in the new boundaries for Colorado were extremely minor. The desire was to retabulate 116th Congress (2019-2021) congressional district profiles for the newly drawn seven districts of Colorado and the eighteen districts of Pennsylvania. The request was to produce these congressional district profiles without using the usual estimation methodology based on American Community Survey (ACS) sample microdata. This was to be done only using data that were already available to the public. For each congressional district, these profiles provide estimates of age distribution, median age, sex (Female/Male) distribution, number of Veterans 18 years and older, race distribution, proportion Hispanic/Latino, distribution of occupied/vacant housing units, and distribution of occupied housing units according to owner occupied and renter occupied.

Highlights: During FY2020, a final study (Wright, Klein, Slud, 2020) was released which documents a methodology and results for retabulating 116th Congress (2019-2021) Congressional District Profiles for the newly drawn eighteen (18) districts of Pennsylvania. The methodology "adjusts" 2010 Census (SF1) population and housing units counts at the block level of old districts to reflect growth from 2010 to 2015 using 2015 ACS 1-year profiles, moves their blocks to the new districts, and retabulates distributions in profiles for new districts with these adjusted counts of housing and persons (population). This work was to be done without using the usual estimation methodology using American Community Survey sample microdata. This work was done only using data that were already available to the public. Inspection of empirical results from the proposed methodology are not too different from those results obtained using the usual methodology. While future work is proposed, for now, this project is complete.

Staff: Tommy Wright (x31702), Martin Klein (FDA), Eric Slud

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

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

Highlights: During FY2020, staff published the Center for Statistical Research and Methodology *(Statistics #2020-02) Study*: "Variability Assessment of Data Treated by the TopDown Algorithm for Redistricting." Unlike an earlier draft where epsilon = 0.25, in our latest work, we take epsilon = 4. From our earlier work, we presented results for other values of epsilon ranging from 0.01 to 10 and observed empirically that variability (over 25 runs of the algorithm) decreases as epsilon increases and that this decrease in variability seems to level off for values of epsilon greater than or equal to 3. The first set of analyses in our latest study is a follow-up to earlier analyses done for Rhode Island in the context of the 2018 End-to-End Census Test. In the second set of analyses in our study, we repeat our analyses for three specific cases provided by the U.S. Department of Justice. As with our earlier work, our approach has two parts: (1) to report observations on variability of results among 25 new runs of the TopDown Algorithm and (2) to report observations on variability between the data results among 25 new runs of the TopDown Algorithm and the published 2010 Census *Public Law 94-171* data. The study focuses on data.

The key empirical message 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).

Staff: Tommy Wright (x31702), Kyle Irimata

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

Highlights: During FY2020, staff explored Bayesian hierarchical models to account for agency-infused noise and spatial dependences in tabulations. Models under consideration included finite mixtures, spatial and Moran's I basis functions, and change-of-support regression. Several approaches were considered to adjust for added DAS noise including a conjugate distribution for Laplace noise, a slice sampler for normally distributed observations with double geometric noise, and a direct sampler based on Walker et al. (*JCGS*, 2011). Staff attended meetings with stakeholders and DAS team to stay informed on larger disclosure avoidance effort. Preliminary results obtained on selected geographies and race/ethnicity groups were communicated to stakeholders.

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

1.9 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 6385F70)

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

Highlights: During FY2020, staff continued work on a project that utilizes CARMA processes to model ACS time series data. A library for the R code was created and documented.

Staff: Tucker McElroy (x33227; R&M), Patrick Joyce

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

Staff: Tommy Wright (x31702), Martin Klein (FDA), Jerzy Wieczorek (Colby College), Nathan Yau

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

Highlights: During FY2020, staff participated in the project planning and corresponding documentation. Staff explored several sources of auxiliary information for use in small area estimation modeling for the predictions supporting the 2021 determinations, including the Citizen Voting Age Population (CVAP) estimates, administrative records, previous vintages of ACS, and decennial data. Staff obtained access to IRS and Numident administrative records, and obtained updated ACS data. Staff cleaned up, commented, and updated SAS code that produces direct estimates and other area-level quantities from the ACS microdata. Staff studied the potential of using temporal small area estimation models that borrow strength from previous vintages of ACS, and developed two new small area temporal models for categorical data that feature a VAR(1) structure in the model errors. Staff also studied analytically the impact of using estimates from the same survey as covariates, while ignoring their error and sampling variance.

Staff: Carolina Franco (x39959), Eric Slud, Xiaoyun Lu, Mark Asiala (DSSD), Tommy Wright

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

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

Highlights: During FY2020, staff implemented two different Bayesian methods to attempt to estimate the number of spatial mixture components that are needed to fit the data. One method is based on the Dirichlet process prior which is computationally intensive. The second method is based on an approximation to the Dirichlet prior process called the 'stick breaking' representation (Sethuraman, 1994) which can be capped at a pre-specified number of mixture components. Computational efficiency of methods was a limiting factor in attempting to ramp up the multivariate spatial mixture model to all counties in the U.S. A reasonable compromise was to fit the state of Minnesota and the surrounding states to borrow nearby information spatially without overwhelming the computational resources. Staff submitted a manuscript titled "Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models with Application to American Community Survey Special Tabulations" to a journal for review.

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

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 FY2020, some additional specifications for the simulation design were drafted, allowing for investigation of small-domain bias in SDR variance estimation, taking into account dependence between domains and stratum differences in outcome variables. Staff in the Demographic Statistical Methods Division coded and proceeded with detailed testing of the augmented simulation programs in R. Results should be tabulated and submitted for publication during FY2021.

Staff: Eric Slud (x34991), Robert Ashmead (The Ohio State University), Tim Trudell (DSMD)

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

A. Data Integration

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

Highlights: During FY2020, staff updated address and name standardizing software for administrative matching support. Staff have archived and are documenting software. Staff defined a new methodology to identify false matches from record linkage, creating a social network of matchings.

Staff: Ned Porter (x31798)

1.12 POPULATION DIVISION PROJECTS (Demographic Project TBA)

A. Introductory Sampling Workshop

Description: In support of Population Division's International Programs Area, staff will conduct (on request) introductory sampling workshops with focus on probability sampling for participants from various countries. These workshops are primarily funded by USAID or other sources.

Highlights: Over the two-week period (January 27-February 7, 2020), staff conducted a Workshop: Introduction to Survey Sampling (focus on Probability Sampling) at the Census Bureau Headquarters. As in the past, the workshop presented the main components of survey sampling with a focus on probability sampling and estimation techniques. Topics included the production of estimates of population parameters from sample surveys as a function of sample design, weighting procedures, the computation of sampling errors of sample estimators, and the making of inferences from the sample to the population. The eight workshop participants were staff members from national statistical agencies: National Population Commission (Nigeria); National Statistical

Office (Malawi); Department of Statistics (Turks and Caicos Islands); Department of Statistics (Bermuda); and the University of the Virgin Islands (Virgin Islands). On the final day, the workshop featured a panel on sampling to give overviews of the American Community Survey, the Business Sample revision Surveys (Retail, Wholesale, Services), and the Current Population Survey.

Staff: Tommy Wright (x31702)

1.13 SOCIAL, ECONOMIC, AND HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165019/7165020)

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 FY2020, staff documented the evaluation of different small area models for estimating the county poverty rate of school-aged children using the 1000 samples from an artificial population. The models are based on the standard Fay-Herriot of the log number of children in poverty and the Binomial Logit Normal model on the county level poverty rates. This evaluation project showed some of the inadequacies of the artificial population and its samples for use at the county level (it was primarily designed to evaluate tract level estimates of poverty). These issues are currently being addressed in the second version of the artificial population project.

Staff also attempted to extend the Dirichlet-Multinomial Share model to add in a component to account for correlation between two sets of shares. For the SAIPE application, these two sets of shares were school district to county share for school-aged children in poverty and school-aged children not in poverty. The added dependence component created a likelihood function that was extremely difficult to evaluate due to the high dimensional non-nested integrals. Several approaches were considered and two were implemented: Laplace approximation and a Monte Carlo integration. Parts of the codebase were rewritten to take advantage of parallel computing resources. Evaluations of the approximations to the likelihood showed that the accuracy was not sufficient for parameter estimation and prediction. Staff also extended the fixed effects part of the share model to include offsets for elementary and secondary school districts (contrasted to unified districts) due to lack of fit that was discovered through graphical assessments of the model estimates. Parts of this project were presented in a talk at the Virtual Joint Statistical Meetings in August 2020.

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

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

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

Highlights: During FY2020, staff prepared a public use dataset for testing based on all person poverty estimates at the county level using 5-year American Community Survey estimates. The predictor variables come from the Population Estimates Program (county population size) and the county-level SNAP (foodstamps) participation counts. Staff has started to read background papers in Geographically Weighted Regressions (GWR) to develop an initial approach to evaluate the homogeneity of parameters across geography. The selection of the bandwidth (distance penalty) is an important part of GWR, and staff is developing a new criteria based on the error of the small area prediction rather than the traditional cross-validation measure.

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

C. Small Area Health Insurance Estimates (SAHIE)
Description: At the request of staff from the Social, Economic, and Housing Statistics Division (SEHSD), our staff will review current methodology for making small area estimates for health insurance coverage by state and poverty level. Staff will work on selected topics of SAHIE estimation methodology, in conjunction with SEHSD.

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

Highlights: During FY2020, staff developed a Bayesian unit-level small area model, which uses a survey-weighted pseudolikelihood, for use with data obtained from an informative survey. A latent Gaussian process model was introduced to account for spatial and multivariate dependencies, and a nonparametric Dirichlet process prior was used to account for clustering in the data. Full conditional distributions, each belonging to standard parametric families, were derived. This model was fit to SAHIE data, with a goal of predicting the number of persons in different IPR categories with health insurance at the county level.

Staff: Ryan Janicki (x35725)

1.14 GENERAL ECONOMIC STATISTICAL SUPPORT (Economic Project 1183X01)

A. Use of Big Data for Retail Sales Estimates

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

Highlights: During FY2020, staff worked on imputation models using establishment-level data from NPD (a commercial retail data aggregator) to produce an experimental data product of retail sales volume by 3-digit NAICS codes and state on a monthly basis. Staff studied a hierarchical Bayesian imputation model for estimating state-level retail sales using establishment-level business register data, state-level economic data and spatial state-level random effects. Center for Statistical Research & Methodology, Economic Indicators Division (EID), and Economic Statistical Methods Division (ESMD) staff developed a blended data product via a composite estimate that combines a synthetic estimate with an estimate based on the mixed effect imputation model. The experimental product – called the Monthly State Retail Sales (MSRS) report – was released and a research paper documenting the imputation model was started. This experimental data product represents a step toward providing more timely, granular, and relevant data products that meet data user needs, while minimizing the burden on respondents. The MSRS is one of the Census Bureau’s first efforts to develop a model that blends traditional survey data with administrative data and third-party data sources, while producing a new data product measuring our rapidly evolving economy.

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

B. Seasonal Adjustment Support

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

Highlights: During FY2020, staff provided seasonal adjustment and software support for users within and outside the Census Bureau, and continued drafting recommendations on assessing residual seasonality in economic time series.

Specific seasonal adjustment and software support for users within and outside the Census Bureau included: INDEC (Argentina), Macrobond, IMT Lucca (Italy), Statistics Norway, Croatia Bureau of Statistics, Government of Japan, Statistics Austria, National Accounts of Peru, INEGI (Mexico), Federal Institute of Technology Switzerland, National Accounts of Columbia, Estima, Nifty, The Trace, Office for National Statistics, Bureau of Labor Statistics, and British Columbia Stats.

Staff provided code and documentation as follows: Ecce Signum to staff at the Office for National Statistics and PricingSolutions; new seasonality diagnostics to the Australian Bureau of Statistics; k-factor GARMA to Jonathan Goldstein of Bowdoin College; VAR parameterization to Sarah Heaps of Newcastle University; mixed frequency time series analysis to Ang Boon Heng of Manpower Research and Statistics Department of Singapore; multi-step ahead testing with Ruey Tsay of University of Chicago.

Staff developed a short document summarizing best modeling practices in view of the Covid-19 crisis, and shared with Richard Penny of Stats New Zealand.

Staff generated and scored simulations for a machine learning seasonal detection methodology created by Gary Cornwall of Bureau of Economic Analysis.

Staff: Tucker McElroy (x33227; R&M), James Livsey, Osbert Pang, William R. Bell (R&M)

C. Seasonal Adjustment Software Development and Evaluation

Description: The goal of this project is a multi-platform computer program for seasonal adjustment, trend estimation, and calendar effect estimation that goes beyond the adjustment capabilities of the Census X-11 and Statistics Canada X-11-ARIMA programs, and provides more effective diagnostics. The goals for FY 2020 include: continuing to develop a version of the X-13ARIMA-SEATS program with accessible output and updated source code so that, when appropriate, the Economic Directorate can produce SEATS adjustments; and incorporating further improvements to the X-13ARIMA-SEATS user interface, output and documentation. In coordination and collaboration with the Time Series and Related Methods Staff of the Economic Statistical Methods Division (ESMD), staff will provide internal and/or external training in the use of X-13ARIMA-SEATS and the associated programs, such as X-13-Graph, when appropriate. Additionally, development efforts are focusing on the future software products that advance beyond current capabilities of X-13ARIMA-SEATS. This new product aims to handling sampling error, treatment of missing values and multivariate analysis. This development is a joint effort with staff from CODS and ESMD.

Highlights: During FY2020, staff from the Center for Statistical Research & Methodology, the Economics Directorate, and the Center for Optimization and Data Science made progress toward modernizing the software maintenance of the X-13ARIMA-SEATS code base. This includes all team members learning Git, Github, and Jira. These software products will allow tracking of bug fixes and new developments for all future modification to the X-13ARIMA-SEATS code base.

This team began to improve the output of the signal extraction with ARIMA time series (SEATS) method within X-13ARIMA-SEATS. The SEATS method was inherited from the Bank of Spain and includes an unusually large amount of default output. This output is being improved and updating the way users control output amount to be consistent with other specs within X-13ARIMA-SEATS.

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

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

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

Highlights: During FY2020, staff made progress on several research projects: (a) continued documentation for the assessment of the impact of weather regressors on seasonal adjustment; (b) obtained new modeling results (allowing for missing values) for the analysis of daily time series (New Zealand immigration data and credit card transaction data); (c) improved methodology for seasonality diagnostics based upon autoregressive roots; (d) formulated several new measures of seasonality, though the statistical methodology needs to be developed; (e) obtained new results on seasonal vector form, useful for understanding the properties of down-sampled time series, which assists the analysis of frequency aggregation; (f) reviewed and documented current practice in detecting residual seasonality, also providing a new asymptotic distribution theory for the QS diagnostics; (g) revised and improved benchmarking optimization methodology for indirect seasonal adjustment of quarterly time series; (h) examined the presence of residual seasonality in large components of GDP using seasonality diagnostics; (i) further developed code and methodology for the use of the EM algorithm in conjunction with signal extraction methods to fit multivariate time series models; (j) refined modeling of weekly and daily time series, including analyses of Business Formation Statistics data and weekly unemployment insurance claims data, with new methods to handle dynamic Box-Cox transforms; and (k) developed a method of extreme-value adjustment based upon stochastic outliers.

Staff: Tucker McElroy (x33227; R&M), James Livsey, Osbert Pang, Thomas Trimbur, William Bell (R&M)

E. Supporting Documentation and Software for Seasonal Adjustment

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

Highlights: During FY2020, support for seasonal adjustment software was provided through a weekly class on Ecce Signum. This class was instructed by Tucker McElroy and attended by people from the Census Bureau and the Bureau of Economic Analysis. The primary presentation material was from the documentation of Ecce Signum which also continues to be improved.

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

F. Redesign of Economic Sample Surveys (Stratification)

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

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

Highlights: During FY2020, staff continued data analyses on potential models in the Economic Directorate sample surveys ACES, ARTS and AWTS for response and outcome variables in terms of potential MOS variables (Total Payroll, total numbers of employees, sales, etc.). As before the goal of these analyses is to develop R functions to express the outcome variance that would be seen in future analyses of these surveys for various potential strata, for application as part of an overall objective function to be used in defining optimal strata. A high-level overview of the research goals and preliminary results was prepared and presented to the Odyssey steering committee. As a consequence, this research will continue and a written report will be delivered to the steering committee in January.

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

G. Exploring New Seasonal Adjustment and Signal Extraction Methods

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

Highlights: During FY2020, research results were presented at the July Economic Area Methodology Seminar (EAMS). This included results comparing X-11 adjustments (the current method in production) to SEATS adjustments. An important component of this comparison was documenting exact SEATS admissible decompositions compared with ARIMA model inputs; a vital point of consideration if SEATS were ever to be used in production.

The team began development of code to evenly compare the two methods under consideration; SEATS and multivariate models via Ecce Signum. This requires custom code to create a bijection between SEATS parameters and the closest analogous Ecce Signum model. This code will ultimately serve to confirm or oppose the utility of multivariate signal extraction for the M3 survey.

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

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

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

In the first study, the investigators initially collected 1,272,000 records of establishments via a grid search on both Yelp and Google Places APIs, based on a combination of geo-coordinates and keywords in the titles of all two-digit NAICS sectors. Records that did not have a website and user reviews are eliminated reducing the collection to approximately 290,000 records. Training and evaluating models for classification purposes require a random sample of business establishments for which their NAICS codes are known. Next, the 290,000 records are then linked to establishments on the Business Register (BR) using the Multiple Algorithm Matching for Better Analytics (MAMBA), a fuzzy matching software developed by Cuffe and Goldschlag. Linkage and other restrictions imposed on selections resulted in a final collection of records for 120,000 single-unit establishments. Employing doc2vec, a text mining technique, the textual information in each record is transformed to vectors. These vectors and series of binary variables indicating the Google Type**,** tags are used as features, i.e. predictors of the NAICS codes. In this approach, Random Forest models are trained to predict NAICS codes. It is reported that the best model performs approximately 59% accurately. Overall, this work initiates an interesting approach for the NAICS classification problem, but one main question is to what extent the methodology can be relied on? One main issue is selection bias and the non-probability nature of the collected data. The data collection process seems to systematically exclude and include business establishments. For example, due to coverage in the source of the collection, grid search selection mechanism and importantly**,** the error-prone record linkage. A closer examination of the data may shed light on these issues.

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

Highlights: During FY2020, different strategies for improving record linkage between the Business Register dataset and the Google scrapped dataset of business establishments were studied. Supervised and unsupervised record linkage in parallel were conducted and their qualities examined.

Staff: Emanuel Ben-David (x37275), Javier Miranda (EWD), Ann Russell (EWD)

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), Joe Engmark, Michael Hawkins, Eric Slud, Kelly Taylor

B. Research Computing

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

Highlights: During FY2020, staff completed the migration of research1 and research2 projects to the Integrated Research Environment (IRE). Final backups of user data for research1 and research2 were completed and these systems were decommissioned. The IRE continues to be scaled up, and now includes forty-one compute nodes. The environment serves both internal and external researchers, including the Federal Statistical Research Data Centers.

Work continued on the Cloud Research Environment (CRE) prototype. The project consists of provisioning computing infrastructure in AWS GovCloud rather than using on-premise servers. For each project that will take place, an environment of machines consisting of a login node and one or more compute nodes are provisioned as Linux EC2-instances. A shared filesystem that can be used by all nodes supporting a project is implemented using Amazon Elastic File System (EFS) volumes mounted via NFS. Four new projects were chosen as candidates for migration to the CRE. These projects include work on formal privacy algorithms for the American Community Survey, record linkage research associated with the 2020 Decennial Census, and supplemental Construction Indicator Research (with the Bureau for Economic Analysis).

Staff: Chad Russell (x33215)

C. Investigative Service Staff Projections Report

Description: Data are provided on U.S. Census Bureau job applicants who had to go through background checks and be cleared by the Office of Personal Management (OPM), U.S. Census Bureau Investigative Services (CIS) and Census Bureau Security staff. The goal of the study is to understand what factors or variables associate with successful background checks and subsequent hiring for employment. Another goal of this study is to determine the time required to clear job applicants and identify models that can accelerate the clearing process.

Highlights: During FY2020, staff performed logistic regression modeling to obtain predictions of the likelihood of completing U.S. Census Bureau job applications, interviews and successful background checks by job applicants. They further conducted model diagnostics to determine the goodness of fit for the logistic regression analyses, and to show the association between observed and predicted outcomes. This work has ended, and a draft report documenting the work has been accomplished.

Staff: Isaac Dompreh (x36801), Kimberly Sellers, Steven Klement (R&M), Mattie Jones (R&M)

D. Survey Experiments to Assess Respondent Preferences and Behavior: Design and Analysis

Description: Experiments are carried out by the Center for Behavioral Science Methods (CBSM) with the goal of collecting data pertaining to respondent preferences and behavior. Both within-subject and between-subject designs have been used in such experiments, including factorial experiments. The data obtained include both binary and continuous responses. The conclusions drawn are used to make recommendations while designing such surveys, for example online surveys using a smartphone.

Highlights: During FY2020, staff provided consulting help to the staff at the Center for Behavioral Science Methods in the analysis of data from several sample survey experiments. These include the following surveys: (i) to investigate the optimal alignment of multi-line text in surveys on smartphone while responding to an online survey, (ii) to explore the effect of different variables on the probability that a respondent would omit someone from the roster section of a decennial census, (iii) to investigate the optimal format for entering USD amounts on the screen accurately and faster, (iv) in a mobile web survey with automated skips, to investigate an optimal method for grouping questions that will lead to best respondent performance when completing a survey, (v) to investigate if respondents carry out a navigation task more accurately, faster, with better satisfaction, with a breadcrumb trail than without, while completing a survey on a smartphone, and (v) to investigate the way to relay interviewer instructions to field enumerators so that the instructions do not distract but rather aid the interviewer in their primary role of obtaining accurate responses to survey questions.

Staff: Thomas Mathew (x35337)

1.16 BUREAU OF ECONOMIC ANALYLSIS (Census Bureau Project TBA)

A. Business Cycle Movements in National Accounts Series

Description: Staff collaborate with researchers from the Bureau of Economic Analysis (BEA) on this project, which aims to analyze business dynamics revealed in Real Gross Domestic Product (GDP) and its major components. It provides an updated set of empirical regularities that are based on recent data and use an adaptive approach that accounts for series' different properties.

Highlights: During FY2020, staff worked with model selection strategies, where a sizable number of candidate forms are available, and where the selection strategy is required to work across diverse data. For Inventory Investment, the series warrants a more pronounced cycle component than consumption service; the quantitative approach allows one to be precise about the differences and to assess the exact impact on filtered trend and cycle series. The key practical issue for economists and policy-makers is how different the results are from the widely used Hodrock-Prescott and Baster-King filters, and this comparison is done in several perspectives. Recently, we have begun to utilize estimated trend and cycle components for forecasting the broadest aggregate of GDP. Our hope is that the modelling method can improve on forecasting this critical series representative of the macroeconomy.

Staff: Thomas Trimbur (x36864), Baoline Chen (BEA)

1.17 FEDERAL HIGHWAY ADMINISTRATION (Census Bureau Project TBA)

A. Modeling and Signal Extraction of Pedestrian and Bicycle Crash Data

Description: The Census Bureau provided statistical expertise to the Federal Highway Administration (FHWA). The research involves modeling and signal extraction of pedestrian and bicycle crash data. In this event level crash data, trends and cycles are obscured by the large amount of seasonality present as well as the sparsity of events at fine levels of geography and time. This collaboration uses Census Bureau demographic data as aggregation factors to amplify the signals then considers signal extraction methodology to account to seasonality.

Highlights: During FY2020, a homogenized dataset of all pedestrian and bicycle crashes in the state of North Carolina was thoroughly vetted and documented. This investigation has resulted in a higher quality data product. Signal extraction and modeling has also been improved.

An investigation into the spatial structure of the data was conducted. This involved fitting a conditional autoregressive (CAR) model to the aggregated county crash counts.

Staff: James Livsey (x33517), Roya Amjadi (FHWA)

1.18 NATIONAL CANCER INSTITUTE

A. National Cancer Center Tobacco Use Survey/Current Population Survey

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

Highlights: During FY2020, staff revised county-level direct survey-based estimates of population coverages of the two tobacco measures of smoke-free workplace policies and smoke-free home rules. County-level estimates for these two tobacco measures were re-calculated for 3,134 counties across the country.

Additionally, model-based estimates for population coverage of smoke-free workplace policies and smoke-free home rules were produced for 3,134 U.S. counties for 2014-15 TUS-CPS files. Bayesian modeling through a Markov Chain Monte Carlo simulation was used to produce the final model-based county-level estimates for our draft manuscript “Small Area Estimation of Smoke-free Workplace Polices and Home Rules in U.S. Counties”. Draft manuscript documenting work accomplished on this project was submitted to the *Journal of Nicotine and Tobacco Research* for publication and is near acceptance.

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

2. RESEARCH

2.1 GENERAL RESEARCH AND SUPPORT (Census Bureau Project 0331000)

Missing Data & Observational Data Modeling

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

Research Problems:

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

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

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

Highlights: During FY2020, staff continued collaboration with an interdivisional team on using nonparametric Bayesian methods developed by Kim et al. (2017) for editing and imputing economic census microdata. We implemented a fully synthetic data generator that integrates editing and multiple imputation processing with data synthesis to produce synthetic microdata that are suitable for sharing with the public. In addition, the team implemented a partially synthetic data generator which produces multiple partially synthetic microdata with edited economic census data as input. We continued work on measuring univariate and multivariate disclosure risk for both fully and partially synthetic data. The disclosure avoidance sub-team established assumptions and developed procedures to assess risk of disclosure in the synthetic data based on isolation of a given multivariate observation and how closely key variables in the synthetic observation are matching the corresponding original multivariate values. The team wrote a research report summarizing our results including recommendation of a suitable synthetic data generator (fully synthetic or partially synthetic data generator) and analyses of disclosure risk vs. utility of the synthetic data.

Staff: Maria Garcia (x31703), Darcy Steeg Morris, Yves Thibaudeau, Jun Shao

B. Imputation and Modeling Using Observational/Alternative Data Sources
Description: This project covers development of statistical methods and models for using alternative source data to supplement and/or replace traditional field data collection. Alternative source data includes administrative records – data collected by governmental agencies in the course of administering a program or service – as well as commercial third party data. Such data often contains a wealth of information relevant to sample surveys and censuses, but suffers from bias

concerns related to, for example, coverage and timeliness. Imputation, classification and general statistical modeling techniques can be useful for extracting good information from potentially biased data.

Highlights: During FY2020, staff worked with Economic Directorate colleagues to develop Bayesian multiple imputation models of state-level retail sales based on data from a third party data aggregator. These imputations are used as input for a more geographically granular and timely estimate than produced by the Monthly Retail Trade Survey (MRTS). With this work as a case study, staff began working with colleagues on a paper illustrating Bayesian multiple imputation hierarchical models in Stan for economic data using third party data.

Staff: Darcy Steeg Morris (x33989), Yves Thibaudeau

Record Linkage & Machine Learning

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

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

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

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

Highlights: During FY2020, a manuscript titled “Two-Stage Approach to Multivariate Linear Regression with Sparsely Mismatched Data” was completed and resubmitted for publication. In this paper, we study the multi-response linear regression with linked data and propose a method for adjusting the analysis to remedy the linkage bias and errors. A two-stage method is proposed under the assumption that the false matches are relatively small. In the first stage, the regression parameter is estimated by handling mismatches as contaminations, and subsequently the correct matches are estimated by a basic variant of sorting. The approach is both computationally convenient and equipped with favorable statistical guarantees. Specifically, it is shown that the conditions for correcting the false matches and recovery become considerably less stringent as the number of responses per observation increase. Numerical results on synthetic and real data are presented to support the main findings of our analysis.

In summary, a new extension to generalized linear models is considered. A new observation-specific method in combination with $\ell 1$-penalization is proposed to account for potential mismatches in more general setting than that of linear regression, and its statistical properties are studied. In this proposal, we give some sufficient conditions for recovering matches in the linked data files using the covariates and responses when the regression parameter is accurately estimated. The proposed approach is compared to established baselines.

Staff: Emanuel Ben-David (x37275)

B. Entity Resolution and Merging Noisy Databases
Description: Work is underway on the problem of merging noisy databases to remove duplicate entities (individuals, households, etc.), where typically a unique

identifier is not known. This problem in the literature is known as entity resolution or record linkage. Work is undertaken on improved methodology, and scalability, and testing such methods on both synthetic and real data.

Highlights: During FY2020, the work of Marchant et al. (2020), which is the first joint Bayesian method to utilize blocking and entity resolution has undergone testing for possible usage for a productionized record linkage methodology. Specifically, the authors have looked at a case study of the decennial census from 2010, where they have merged this with administrative records for the state of Wyoming. In addition, the authors along with Center for Optimization & Data Science staff are continuing to test this on additional states to see its applicability for productionized software moving forward. Marchant et al. (2020) has been accepted in *the Journal of Computational and Graphical Statistics.*

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

Small Area Estimation

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

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

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

A. Using ACS Estimates to Improve Estimates from Smaller Surveys via Bivariate Small Area Estimation Models

Description: Staff will investigate the use of bivariate area-level models to improve small area estimates from one survey by borrowing strength from related estimates from a larger survey. In particular, staff will explore the potential of borrowing strength from estimates from the American Community Survey, the largest U.S. household survey, to improve estimates from smaller U.S. surveys, such as the National Health Interview Survey (NHIS), the Survey of Income and Program Participation, and the Current Population Survey.

Highlights: During FY2020, staff created user-friendly code that illustrates how each of the three proposed bivariate models can be run in practice using R and JAGS. Staff finalized an associated paper. Staff performed revisions to the paper which was subsequently accepted to the *Journal of Survey Statistics and Methodology*.

Staff: Carolina Franco (x39959), William R. Bell (R&M)

B. 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 FY2020, staff explored a bootstrapping approach free of distributional assumptions to estimate the MSE for small area models. Staff utilized a bootstrapping distribution matching the variance and kurtosis of the random effects. Staff conducted simulation studies and observed lower bias for the bootstrap approach as compared to traditional estimators of MSE. This discrepancy is of particular note when the sampling variances are large as compared to the random effects variance, as can be the case in practice. Staff presented initial results at the 2020 Joint Statistical Meetings, and a manuscript is in preparation to document findings.

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

C. Small Area Estimation for Misspecified Models

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

Highlights: It is well known that the usefulness of an EBLUP of a small area mean depends on the correctness of the assumed model. In particular, approximation and traditional estimators of MSE have not been evaluated under misspecified models. One particular form of model misspecification is a failure to include a useful covariate in the mean model. During FY2020 and in an ongoing project, staff is collaborating with Professors Snigdhansu Chatterjee and Abhyuday Mandal to investigate the impact of a missing covariate on the MSE estimation. Our preliminary findings show that the traditional second-order unbiased MSE estimators substantially underestimate the true MSE if the assumed model fails to include a useful covariate. Such omission of even a "marginally influential regression coefficient" invalidates the second-order MSE estimation results. These results are based on method of moments estimation of model parameters.

Staff is conducting simulations to evaluate impact of a missing covariate on the MSE of the misspecified EBLUP. Staff is also exploring other estimating equation based method of estimating model parameters.

Staff: Gauri Datta (x33426), Eric Slud

D. 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 FY2020, staff and external collaborators completed a manuscript exploring effectiveness of various spatial random effects models as alternative to the Fay-Herriot model. We assess the effectiveness of these spatial models based on a simulation study and a real application. We consider the prediction of statewide four-person family median incomes for the U.S. states based on the 1990 Current Population Survey and the 1980 Census. This application and simulation study show considerably superior performance of some of the spatial models over the regular Fay-Herriot model when good covariates remain unavailable. In some applications, some small areas are created after completion of a survey that does not provide any direct estimates of the late-breaking unsampled small areas. Proposed spatial models generate better predictions of unsampled small area means by borrowing from neighboring residuals than the synthetic regression means that result from regular independent random effects Fay-Herriot model. Datta presented this work in a CSRM seminar in July.

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

Sampling Estimation & Survey Inference

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

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

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

Research Problems:

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

Potential Applications:

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

A. Household Survey Design and Estimation

[See Demographic and ACS Projects]

B. Sampling and Estimation Methodology: Economic Surveys

Description: The Economic Directorate of the Census Bureau encounters a number of issues in sampling and estimation in which changes might increase the accuracy or efficiency of the survey estimates. These include estimates of low-valued exports not currently reported, alternative estimation for the *Quarterly Financial Report*, and procedures to address nonresponse and reduce respondent burden in the surveys. Further, general simulation software might be created and structured to eliminate various individual research efforts. An observation is considered influential if the estimate of total monthly revenue is dominated by its weighted contribution. The goal of the research is to find methodology that uses the observation but in a manner that assures its contribution does not dominate the estimated total or the estimates of period-to-period change.

Highlights: During FY2020, there was no significant progress.

Staff: Mary Mulry (682-305-8809)

C. The Ranking Project: Methodology Development and Evaluation

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

Highlights: During FY2020, staff published theory (Klein, Wright, Wieczorek, 2020) for a simple and useful joint confidence region for an overall ranking of K populations that gives a measure of uncertainty for the estimated overall ranking based on estimates from sample data. A proposed visualization makes it easy to communicate this uncertainty in the estimated overall ranking while also revealing many other possible overall rankings. Unlike previous research which tends to focus on uncertainty in the individual ranks, this focuses on uncertainty in the overall ranking. The theory assumes that we are able to provide known upper and lower bounds for the parameter used for ranking each population. The theory is developed in a frequentist setting, but we indicate how the methodology can be adapted for a Bayesian setting. A little work has begun to look at associated visualizations and to optimize or make the joint confidence region tighter.

Staff: Tommy Wright (x31702), Martin Klein (FDA), Jerzy Wieczorek (Colby College), Nathan Yau

D. Sampling and Apportionment

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

Sample Allocation

Highlights: During FY2020, staff published research results (Wright, 2020) on a general exact optimal sample allocation algorithm, with bounded cost and bounded stratum sample sizes. The research shows that some known allocation methods are special cases.

Staff: Tommy Wright (x31702)

Apportionment

Highlights: During FY2020, there was no significant progress.

Staff: Tommy Wright (x31702)

E. Consistent Estimation of Mixed-Effect Superpopulation-Model Parameters in Complex Surveys with Informative Sampling

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

Highlights: During FY2020 and at a 2020 Joint Statistical Meetings (JSM) session, staff delivered the talk and prepared and submitted a JSM proceedings paper describing and slightly extending the research that went into the talk. The major outcome was a mathematical proof showing that the mixed-effect variance parameters in an informatively sampled two-level superpopulation are not in principle identifiable from design-based estimators making use only of first-order inclusion probabilities from the survey design.

Staff: Eric Slud (x34991)

F. Comparison of Probability (RDD) and Nonprobability in a Census Tracking System

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

Highlights: During FY2020, work with the Center for Behavioral Science Methods Tracking Survey continued. The main effort was a continuing investigation of discrepancies between either weighted or unweighted Telephone Survey results and national-survey benchmarks to try to discover in what way deficiencies in the RDD data-collection might have caused these discrepancies. As part of this effort, staff drafted a write-up of base-weighting methodology taking account of missing data in the telephone-type and Household-size data used in the contractor's base-weighting.

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

Time Series & Seasonal Adjustment

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

Research Problems:

• All contemporary seasonal adjustment programs of interest depend heavily on time series models for trading day and calendar effect estimation, for modeling abrupt changes in the trend, for providing required forecasts, and, in some cases, for the seasonal adjustment calculations. Better methods are needed for automatic model selection, for detection of inadequate models, and for assessing the uncertainty in modeling results due to model selection, outlier identification and non-normality. Also, new models are needed for complex holiday and calendar effects.

• Better diagnostics and measures of estimation and adjustment quality are needed, especially for model-based seasonal adjustment.

• For the seasonal, trading day and holiday adjustment of short time series, meaning series of length five years or less, more research into the properties of methods usually used for longer series, and perhaps into new methods, are needed.

Potential Applications:

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

A. Seasonal Adjustment

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

Highlights: During FY2020, staff made progress on several research projects: (a) continued work on reconciliation for aggregates of time series, with extensions to the conversion of monthly to quarterly flow time series, such that seasonal adjustment adequacy is preserved. The methodology was updated by a new optimization strategy that speeds up the implementation, and with a superior initialization; (b) continued work on signal extraction diagnostics methodology, including new diagnostics for seasonal adjustment based on autoregressive roots. Simulation studies were completed and theory developed, with extensions to the case of seasonal unit roots and the problem of over-adjustment. Also, a Wald statistic was developed for faster computation, with a different variance estimate that allows for model mis-specification; (c) continued refining the Expectation-Maximization algorithm to assist with the fitting of multivariate time series models as well as the calculation of signal extraction estimates. The methods have been coded and tested on low-dimensional time series; (d) examined the sampling error contribution to seasonal adjustment mean squared error; (e) continued writing code and text for a book on multivariate real-time seasonal adjustment and forecasting. Recent work includes extensions to co-integrated processes and mixed-frequency data; (f) completed new algorithmic work on multivariate seasonal adjustment and missing value imputation. A new software product (Ecce Signum) has been developed to implement the missing value methods, allowing for a broad range of applications in forecasting and extreme-value adjustment; and (g) obtained new formulas for midcasting of nonstationary stochastic processes, allowing for appropriate initial value assumptions, which is needed in the Gibbs sampling routine of an extreme-value adjustment procedure.

Staff: Tucker McElroy (x33227; R&M), James Livsey, Osbert Pang, Anindya Roy

B. Time Series Analysis

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

Highlights: During FY2020, staff made progress on several projects: (a) additional simulations were completed to revise a paper that describes estimation of multivariate time series models using the Frobenius norm; (b) developed a method of model identification based on testing for zeroes in nonparametric estimates of the spectral density; (c) updated simulations for a paper that models the behavior of forecasters over multiple horizons; (d) continued theoretical and applied work on nonlinear prediction for time series forecasting, using Hermite polynomials. A facet of this work involves using autocumulants to construct optimal quadratic predictors. New mathematical techniques were developed for analysis and inversion of 2- and 4-dimensional covariance arrays with Toeplitz structure, and a new factorization result for polyspectra was obtained; (e) updated the code and simulation results for a method to estimate multivariate inverse autocovariances, which is useful for fitting vector moving averages; (f) continued research into methods for count time series, including multivariate modeling that allows for sparse parameterizations; (g) continued research into new models for business cycles, and applied these concepts to components of GDP, with an interest in how much cyclical relationships have been altered by the Great Recession, and also how an estimation strategy that adapts to each series' dynamics improves over ad hoc filtering; (h) developed new methodology for modeling outlier processes, which provides an alternative approach to extreme value adjustment. New sampling algorithms (and full conditional calculations) were developed to enhance the computation; (i) developed code and methodology for local spectral density estimation that is optimal at the boundary of the frequency domain. Theoretical results on optimal bandwidth selection were derived and tested; (j) developed a random permutation modification of "skip-sampling" methodology, a sort of time series bootstrap based on scrambling the discrete Fourier Transform; (k) derived new polyspectral density estimators, and estimators of polyspectral means, a type of weighted integral of polyspectra. New asymptotic results were developed for the polyspectral means; and (l) devised a new idea on confidence intervals for forecasts from non-linear processes.

Staff: Tucker McElroy (x33227; R&M), James Livsey, Osbert Pang, Anindya Roy, Thomas Trimbur

C. Time Series Model Development

Description: This work develops a flexible integer-valued autoregressive (AR) model for count data that contain data over- or under-dispersion (i.e. count data where the variance is larger or smaller than the mean, respectively). This model contains Poisson and negative binomial AR models as special cases.

Highlights: During FY2020, research was published; project is complete.

Staff: Kimberly Sellers (x39808)

Experimentation, Prediction, & Modeling

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

Research Problems:

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

Potential Applications:

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

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

Highlights: During FY2020, (1) Staff developed a generalization of the multinomial distribution extended via the CMP distribution, thus establishing a Conway-Maxwell-multinomial (CMM) distribution. Staff published the research in the *Journal of Multivariate Analysis* and created the COMMultReg R package to demonstrate methods used in the manuscript. (2) Staff are developing various constructions of a multivariate CMP distribution and studying their respective properties and related statistical matters. (3) Staff developed a first-order moving average model based on the sum-of-CMPs distribution which contains the first-order Poisson and negative binomial moving average models as special cases.

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

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

Highlights: During FY2020, staff completed a paper on sample size determination under fixed effects Continuation-Ratio Logit model, which is also mentioned under Project E. "Experiment for Effectiveness of Bilingual Training" in Section 1 of the report.

Staff: Andrew Raim (x37894), Thomas Mathew, Kimberly Sellers

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

Highlights: During FY2020, we considered algorithms capable of implementing large scaled Generalized Additive Models with interactions via L0 regularization. We also explored structured sparsity for more meaningful interpretability. Experiments were tested on the Census Planning Database (PDB). A draft of a working paper is under preparation.

Staff: Emanuel Ben-David (x37275), Shibal Ibrahim (MIT), Rahul Mazumder (MIT)

Simulation, Data Science, & Visualization

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

Research Problems:

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

Potential Applications:

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

A. Development and Evaluation of Methodology for Statistical Disclosure Control

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

Highlights: During FY2020, staff continued work on developing theory and methods for protecting privacy and data confidentiality. Staff published a paper titled "Post-randomization for Controlling Identification Risk in Releasing Microdata from General Surveys" in the *Journal of Applied Statistics*. This paper presents a method for controlling identification risks while well preserving the weighted estimates from a general survey. Staff wrote a technical report titled "A Local l-Diversity Mechanism for Privacy Protected Categorical Data Collection." It presents a new randomized response scheme that is easy to use and offers a better trade-off between privacy protection and statistical efficiency than an optimal procedure under local differential privacy.

Staff: Tapan Nayak (x35191)

B. Bayesian Analysis of Singly Imputed Synthetic Data

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

Highlights: During FY2017-FY2020, staff members (Martin Klein and Bimal Sinha) developed exact inference procedures for singly imputed synthetic data analysis under a frequentist paradigm. This was done for a number of parametric probability models and main emphasis was in the context of a multiple linear regression model with applications to ACS data. Under the current proposal, staff members will conduct Bayesian analysis of singly imputed synthetic data. This is currently under preparation. A draft technical report is now complete.

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

C. Frequentist and Bayesian Analysis of Multiply Imputed Synthetic Data

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

Highlights: During 2017-2020, staff members (Martin Klein and Bimal Sinha) developed exact inference procedures for multiply imputed synthetic data analysis under a frequentist paradigm. This was done for a number of parametric probability models and main emphasis was in the context of a multiple linear regression model with applications to ACS data. Under the current proposal, staff members will conduct a refined frequentist analysis and Bayesian analysis of multiply imputed synthetic data. The research on this topic is now complete, and a draft technical report is being prepared.

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

D. Statistical Meta-Analysis-Local Power Comparison of Several Exact Tests for a Common Mean of Independent Univariate Normals with Unequal Variances

Description: Statistical Meta-Analysis - combining independent tests for a common parameter based on results from several independent studies can arise in many instances. In particular, in the context of a common normal mean with unequal variances from different sources, there is a huge literature suggesting many approximate and exact tests. Under this research project, performances of seven (7) exact tests in terms of local power will be carried out.

Highlights: During FY2020, expressions for local powers of seven (7) exact tests were derived and compared. The investigation is complete, and a *CSRM Research Report* has been prepared and has also been accepted for external publication.

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

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

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

Highlights: During FY2020, there was no significant progress reported.

Staff: Bimal Sinha (x34890)

Summer at Census

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

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

Highlights: During FY2020, nominations were received in response to a call for the *2020 SUMMER AT CENSUS Program.* However, for health and safety reasons due to the Coronavirus (Covid-19 Pandemic), *2020 SUMMER AT CENSUS* was cancelled.

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: Joe Engmark, Michael Hawkins, Kelly Taylor

3. PUBLICATIONS

3.1 JOURNAL ARTICLES, PUBLICATIONS

Baker, S., McElroy, T., and Sheng, X. (2020). “Expectation Formation Following Large and Unpredictable Shocks,” *Review of Economics and Statistics 102 (2),* 287-303.

Barseghyan, L., Molinari, F., Morris, D.S., and Teitelbaum, J. (2020). “The Cost of Legal Restrictions on Experience Rating,” *Journal of Empirical Legal Studies, 17(2),* 38-70.

Ben-David, E. and Rajaratnam, B. (2020). “On the Letac-Massam Conjecture and Existence of High Dimensional Bayes Estimators for Graphical Models,” *Electronic Journal of Statistics, Volume 14, Number 1 (2020), 580-604.*

Binette, O. and Steorts, R. (2020). Discussion of “Multiple-Systems Analysis for the Quantification of Modern Slavery: Classical and Bayesian Approaches,” *Journal of the Royal Statistical Society, Series A*.

Enamorado, T. and Steorts, R. (In Press). “Probabilistic Blocking and Distributed Bayesian Entity Resolution,” *Privacy in Statistical Databases (Lecture Notes in Computer Science* 12276), ed. Josep Domingo-Ferrer and Krishnamurty Muralidhar, 224-239; \href{https://link.springer.com/book/10.1007/978-3-030-57521-2}{https://doi.org/10.1007/978-3-030-57521-2$_$16}.

Edirisinghe, P., Mathew, T., and Peiris, T.S.G (2020). “Confidence Limits for Compliance Testing Using Mixed Acceptance Criteria,” *Quality and Reliability Engineering International*, 36, 1197-1204.

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

Ghosh, M., Kubokawa, T., and Datta, G. (2020). “Density Prediction and the Stein Phenomenon,” *Sankhya, A, 82,* 330-352.

Goyal, S., Datta, G., and Mandal, A. (2020). “A Hierarchical Bayes Unit-Level Small Area Estimation Model for Normal Mixture Populations,” *Sankhya, B.*

Jia, Y., Lund, R., and Livsey, J. (2019). “Superpositional Stationary Count Time Series,” *Probability in the Engineering and Informational Sciences*, 1-19, doi:10.1017/50269964819000433.

Klein, M., Wright, T., and Wieczorek, J. (2020). “A Joint Confidence Region for an Overall Ranking of Populations,” *Journal of the Royal Statistical Society, Series C, 69, Part 3, 589-606.*

Lin, W., Huang, J., and McElroy, T. (2020). “Time Series Seasonal Adjustment Using Regularized Singular Value Decomposition,” *Journal of Business and Economics Statistics 38(3)*, 487-501.

Marchant, N., Kaplan, A., Rubenstein, B., Elzar, D., and Steorts, R.C. (In Press). “d-blink: Distributed End-to-End Bayesian Entity Resolution,” *Journal of Computational Graphics and Statistics*.

McElroy, T. (In Press). “A Diagnostic for Seasonality Based Upon Polynomial Roots of ARMA Models,” *Journal of Official Statistics*.

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

McElroy, T. and Wildi, M. (2020). “Multivariate Direct Filter Analysis for Real-Time Signal Extraction Problems,” *Econometrics and Statistics*, 14, 112-130.

Morris, D. S., Raim, A. M., and Sellers, K. F. (2020). “A Conway–Maxwell-Multinomial Distribution for Flexible Modeling of Clustered Categorical Data,” *Journal of Multivariate Analysis*, vol. 179. https://doi.org/10.1016/j.jmva.2020.104651.

Mulry, M., Bates, N., and Virgile, M. (2020). “Viewing Participation in Censuses and Surveys through the Lens of Lifestyle Segments.” *Journal of Survey Statistics and Methodology*. DOI: 10.1093/jssam/smaa006

Parker, P. A., Holan, S. H., and Janicki, R. (In Press). “Bayesian Unit-Level Modeling of Count Data under Informative Sampling Designs,” *Stat*.

Raim, A.M., Holan, S.H., Bradley, J.R., and Wikle, C.K. (In Press). “Spatio-temporal Change of Support Modeling with R,” *Computational Statistics*. https://doi.org/10.1007/s00180-020-01029-4

Sellers, K.F., Peng, S.J., and Arab, A. (2020). “A Flexible Univariate Autoregressive Time-Series Model for Dispersed Count Data,” *Journal of Time Series Analysis*, 41 (3): 436-453.

Slawski, M., Ben-David, E., and Li, P. (2020). “Two-Stage Approach to Multivariate Linear Regression with Sparsely Mismatched Data,” *Journal of Machine Learning Research*, *21*, 1-42.

Steorts, R., Schmid, T., and Tzavdis, N. (In Press). “Smoothing and Benchmarking for Small Area Estimation with Application to Rental Prices in Berlin,” *International Statistical Review*.

Tancredi, A., Steorts, R., and Liseo, B. (2020). “A Unified Framework for De-Duplication and Population Size Estimation,” *Bayesian Analysis, 15:2,* 633-682.

Tang, J., Reiter, J., and Steorts, R. (In Press). “Bayesian Modeling for Simultaneous Regression and Record Linkage,” *Privacy in Statistical Databases (Lecture Notes in Computer Science).*

Woody, J., Lu, Q., and Livsey, J. (2020). “Statistical Methods for Forecasting Daily Snow Depths and Assessing Trends in Inter-Annual Snow Depth Dynamics,” *Environmental and Ecological Statistics, 27, 3,* 609-628.

Wright, T. (In Press). “A General Exact Optimal Sample Allocation Algorithm: With Bounded Cost and Bounded Sample Sizes,” *Statistics and Probability Letters, 165.*

Wright, T. (In Press). “From Cauchy-Schwartz to the U.S. House of Representatives: Applications of Lagrange’s Identity,” *Mathematics Magazine*.

Yao, X. and Slud, E. (2019). “Nonexistence of an Unbiased Estimation Function for the Cox Model,” *Statistics, and Probability Letters*, 152, 122-127.

Young, D. and Mathew, T. (2020). “Nonparametric Hyper-rectangular Tolerance and Prediction Regions for Setting Multivariate Reference Regions in Laboratory Medicine,” *Statistical Methods in Medical Research*, *29,* 3569-3585.

Zhang, C. and Nayak, T.K. (2020). “Post-Randomization for Controlling Identification Risk in Releasing Microdata from General Surveys,” *Journal of Applied Statistics*, DOI: 10.1080/02664763.2020.1732310.

3.2 BOOKS/BOOK CHAPTERS

Chung, H.C., Datta, G.S., and Maples, J. (2019). “Estimation of Median Incomes of the American States: Bayesian Estimation of Means of Subpopulations,” In "*Opportunities and Challenges in Development Essays*" for Sarmila Banerjee. Eds.: S. Bandyopadhyay and M. Dutta, Springer, 505-518.

Ericuilescu, A., Franco, C., and Lahiri, P. (In Press). “Use of Administrative Records in Small Area Estimation,” in A.Y. Chung and M. Larsen (Eds.), *Administrative Records for Survey Methodology*, New York, NY: Wiley Publishers.

McElroy, T. and Politis, D. (2020). *Time Series: A First Course with Bootstrap Starter.* New York: Chapman Hill.

Nichols, E., Olmsted-Hawala, E., Raim, A., and Wang, L. (2020). “Attitudinal and Behavioral Differences between Older and Younger Adults Using Mobile Devices” in Q. Gao and J. Zhou (Eds.), *Human Aspects of IT for the Aged Population: Technologies, Design, and the User Experience*, Springer Nature Switzerland AG, 325-337.

3.3 PROCEEDINGS PAPERS

Joint Statistical Meetings, American Statistical Association, Denver, Colorado, July 27-August 1, 2019
2019 Proceedings of the American Statistical Association

- Carolina Franco and William Bell, “Using American Community Survey Data to Improve Estimates from Smaller Surveys through Bivariate Small Area Estimation Models.”
- Jerry Maples, “Small Area Estimates of the Child Population and Poverty in School Districts Using Dirichlet-Multinomial Models.”
- Mary Mulry, Yazmin Trejo, and Nancy Bates, “Variable Selection for Multinomial Logistic Regression Modeling to Assign One of Six Census Mindsets Using Big Data.”
- Eric Slud, “Model-Assisted Estimation of Mixed-Effect Model Parameters in Complex Surveys.”
- Tommy Wright, “An Elementary Derivation of Kadane's Optimal Dynamic Sampling Plan.”
- Xiaouu Zhai and Tapan Nayak, “Identity Disclosure Control in Microdata Release by Post-Randomization.”

3.4 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY RESEARCH REPORTS

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

RR (Statistics #2019-08): Tapan K. Nayak and Xiaoyu Zhai, “A Local I-Diversity Mechanism for Privacy Protected Categorical Data Collection,” October 22, 2019.

RR (Statistics #2019-09): Carolina Franco and William R. Bell, “Using American Community Survey Data to Improve Estimates from Smaller Surveys through Bivariate Small Area Estimation Models,” October 25, 2019.

RR (Statistics #2020-01): Xiaoyu Zhai and Tapan K. Nayak, “A Post-randomization Method for Rigorous Identification Risk Control in Releasing Microdata,” April 17, 2020.

RR (Statistics #2020-02): Yehenew G. Kifle, Alain M. Moluh, and Bimal K. Sinha, “Comparison of Local Powers of Some Exact Tests for a Common Normal Mean with Unequal Variances,” May 20, 2020.

RR (Statistics #2020-03): Andrew M. Raim, Thomas Mathew, Kimberly F. Sellers, Renee Ellis, and Mikelyn Meyers, “Experiments on Nonresponse Using Sequential Regression Models,” August 17, 2020.

RR (Statistics #2020-04): Carolina Franco, “Comparison of Small Area Models for Estimation of U.S. County Poverty Rates of School Aged Children Using an Artificial Population and a Design-Based Simulation,” August 24, 2020.

RR (Statistics #2020-05): Eric V. Slud, “Model-Assisted Estimation of Mixed-Effect Model Parameters in Complex Surveys,” August 28, 2020.

3.5 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY STUDY SERIES

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

SS (Statistics #2020-01): Tommy Wright, Martin Klein, and Eric Slud. “A Deterministic Retabulation of Pennsylvania Congressional District Profiles from 115th Congress to 116th Congress,” March 30, 2020.

SS (Statistics #2020-02): Tommy Wright and Kyle Irimata, “Variability Assessment of Data Treated by the TopDown Algorithm for Redistricting,” September 4, 2020.

4. TALKS AND PRESENTATIONS

Mathematics Department Seminar, Morgan State University, Baltimore, Maryland, October 9, 2019.

- Tommy Wright, “Lagrange's Identity and Apportionment of the U.S. House of Representatives.”

International Conference on Statistical Distributions and Applications, Grand Rapids, Michigan, October 10, 2019.

- Kimberly Sellers, “A Flexible Univariate Autoregressive Time-Series Model for Dispersed Count Data.”

Data Linkage Day 2019, Committee on National Statistics, National Academies, Washington, D.C., October 18, 2019.

- Emanuel Ben-David, “Linear Regression with Linked Data Files.”
- Edward Porter, Poster Session, “False Duplicates in the Census: A Novel Approach for Identifying False Matches from Record Linkage Software.”

Bureau of Economic Analysis Seminar, Washington, D.C., October 22, 2019.

- Tucker McElroy, “Assessing Residual Seasonality in the U.S. National Income and Product Accounts Aggregates.”

Mathematics Department Colloquium Series, U.S. Naval Academy, Annapolis, Maryland, October 30, 2019.

- Tommy Wright, “Lagrange's Identity and Apportionment of the U.S. House of Representatives.”

Central American Monetary Council, San Jose, CA, November 4-8, 2019.

- James Livsey and Demetra Lytras, Week long course teaching: “Seasonal Adjustment with X-13ARIMA-SEATS.”

University of Maryland at College Park, Department of Mathematics, College Park, Maryland, November 14, 2019.

- Carolina Franco, Speaker, Alumni Series, University of Maryland Chapter of Women in Math.

North Carolina State University Statistics Colloquium, Raleigh, North Carolina, November 22, 2019.

- Eric Slud, “Estimating a Two-Way Random Effects Model from Informatively Sampled Survey Data.”

International Conference on Environmental and Medical Statistics, University of Peradeniya, Sri Lanka, January 9-10, 2020.

- Tapan Nayak, Keynote, “Statistical Methods for Privacy and Data Confidentiality Protection in Medical Research.”

University of Maryland Statistics Seminar, University of Maryland at College Park, College Park, Maryland, February 27, 2020.

- Eric Slud, “Estimating a Two-Way Random Effects Model from Informatively Sampled Survey Data.”

2020 American Statistical Association Florida Chapter Annual Meeting, University of West Florida, Pensacola, Florida, March 6-7, 2020.

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

ENAR 2020 Spring Meeting, Nashville, Tennessee, March 22-25, 2020.

- Rebecca C. Steorts, “Distributed Bayesian Entity Resolution.”

Symposium on Data Science and Statistics (SDSS), Virtual Symposium, June 3-5, 2020.

- Rebecca C. Steorts, “Challenges of Working with Administrative and Census Data.”

Summer Program in Research and Learning (SPIRAL) and Summer Program Advancing Techniques in the Applied Learning of Statistics (SPATIAL-Stats) Colloquium (virtual), American University, Washington, D.C., June 25, 2020.

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

Joint Statistical Meetings, American Statistical Association, (Virtual, USA), August 2-6, 2020.

- Gauri Datta, “A Hierarchical Bayes Unit-Level Small Area Estimation Model for Normal Mixture Populations.”
- Carolina Franco, Roundtable discussion leader, “Confidence Intervals for Proportions in Complex Sample

Surveys."

- Carolina Franco, Discussant, Combining Information from Multiple Sources for Various Purposes (Session).
- Kyle Irimata, Gauri Datta, and Jerry Maples, "Comparative Study of MSE Estimates for Small Area Models under Different Sampling Variances."
- Ryan Janicki, "Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models with Application to American Community Survey Special Tabulations."
- Jerry Maples, "A Bivariate Dirichlet-Multinomial Small Area Share Model with Application to Joint Estimation of School District Population and Poverty."
- Darcy Morris, Andrew Raim, and Kimberly Sellers, "Conway-Maxwell-Multinomial Regression for Categorical Data with Associated Trials."
- Mary Mulry, Steven Scheid, and Darcy Morris, "Evaluation of Multi-Class Classification Models for Census Mindsets."
- Osbert Pang and William Bell, "Assessing the Contribution of Sampling Variance to Seasonal Adjustment Mean Squared Error."
- Andrew Raim, Thomas Mathew, and Kimberly Sellers, "Sample Size Selection in Continuation-Ratio Logit Mixed Effects Models."
- Eric Slud, "Consistent Mixed-Model Parameter Estimation Under Informative Sampling Using Only Single-Inclusion Weights."
- Rebecca C. Steorts, Neil Marchant, Ben Rubinstein, Daniel Elzar, and Andee Kaplan, "d-blink: Distributed End-to-End Bayesian Entity Resolution."
- Thomas Trimbur and William Bell, "Testing for Time-Variation in Trading-Day Effects on Monthly Time Series."
- Tommy Wright, Panelist, Models for Mentoring (Session).

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

Bradley Efron, Stanford University, Special Video of International Prize in Statistics - "Prediction, Estimation, and Attribution," November 19, 2019.

Joe Schafer, U.S. Bureau of the Census, "Estimating Citizen Voting Age Population in 2020: A Latent-Class Approach," December 16, 2019.

Terrance D. Savitsky, U.S. Bureau of Labor Statistics, Matthew R. Williams, National Science Foundation, Jingchen Hu, Vassar College, “Bayesian Pseudo Posterior Mechanism under Differential Privacy,” January 7, 2020.

Gauri Datta, University of Georgia/U.S. Bureau of the Census, “Spatial Random Effects Small Area Models,” July 8, 2020.

Bimal Sinha, University of Maryland, Baltimore County/U.S. Bureau of the Census, “Comparison of Local Powers of Some Exact Tests for a Common Normal Mean with Unequal Variances,” August 19, 2020.

Adam Hall, U.S. Bureau of the Census Postdoctoral Fellow, “A Unified Price Index for Spatial Comparisons,” September 1, 2020.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

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

- **Michael Ikeda** – 2018 End-to-End Test Disclosure Team "For successful execution of the 2018 End-to-End Test Disclosure Avoidance System that generated microdata in a formerly private manner, while satisfying complex requirements, thus demonstrating the feasibility of utilizing high-quality and rigorous disclosure avoidance protection to be applied to the 2020 Decennial Census."

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

- **Brett Moran** – 2018 End-to-End Test Disclosure Team "For successful execution of the 2018 End-to-End Test Disclosure Avoidance System that generated microdata in a formerly private manner, while satisfying complex requirements, thus demonstrating the feasibility of utilizing high-quality and rigorous disclosure avoidance protection to be applied to the 2020 Decennial Census."

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

- **Edward Porter** – 2018 End-to-End Test Disclosure Team "For successful execution of the 2018 End-to-End Test Disclosure Avoidance System that generated microdata in a formerly private manner, while satisfying complex requirements, thus demonstrating the feasibility of utilizing high-quality and rigorous disclosure avoidance protection to be applied to the 2020 Decennial Census."

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

- **Chad Russell** – Research Data Center Migration to Integrated Research Environment (IRE) Team "The Research Data Center migration to the Integrated Research Environment (IRE) allows researchers access to a single repository for data sharing and collaboration. It combines internal Census Bureau data with external data accessed by university researchers at Federal Statistical Research Data Centers. IRE benefits 1,000 researchers, including 300 Census Bureau and 700 external researchers."

6.2 SIGNIFICANT SERVICE TO PROFESSION

Emanuel Ben-David

- Reviewed papers for *Mathematical Reviews*, 2020 International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2020)
- Refereed a paper for *Survey Methodology*
- Organizer, Invited Session on Record Linkage, 2020 Joint Statistical Meetings
- Member, Ph.D. Defense Committee, Department of Statistics, George Mason University
- Member, Ph.D. Defense Committee, Department of Mathematics and Statistics, University of Maryland, Baltimore County

Gauri Datta

- Associate Editor, *Sankhya*
- Guest Co-Editors, *Sankhya*, Special Issue (Memorial volume for J.K. Ghosh)
- Associate Editor, *Statistical Methods and Applications*
- Associate Editor, *Environmental and Ecological Statistics*
- Editorial Member, *Calcutta Statistical Association Bulletin*

Carolina Franco

- Associate Editor, *Journal of the Royal Statistical Society-Series A*
- Chair, ASA's COWIS Gertrude Cox Scholarship Subcommittee
- Vice-Chair, ASA's Committee on International Relations in Statistics
- Session Chair, Data Integration in 21st Century Government Surveys, 2020 Joint Statistical Meetings

Kyle Irimata

- Refereed papers for *Survey Methodology* and *Statistical Methods in Medical Research*

Patrick Joyce

- Refereed a paper for *Journal of Official Statistics*

Xiaoyun Lu

- Reviewer, *Mathematical Reviews*

Jerry Maples

- Refereed papers for *Journal of Official Statistics, Journal of the American Statistical Association,* and *Survey Methodology*

Thomas Mathew

- Associate Editor, *Sankhya*
- Associate Editor, *Journal of Multivariate Analysis*

Tucker McElroy

- Reviewed submissions for the Zellner Thesis Award.
- Refereed papers for the *Acta Scientiarum Mathematicarum*, *Journal of Applied Econometrics, Journal of Business and Economics Statistics, Journal of Econometrics, Journal of Official Statistics, Journal of Time Series Analysis, Statistical Theory and Related Fields, Econometrics and Statistics, Econometric Reviews, and Quarterly Review of Economics and Finance*

Darcy Morris

- Associate Editor, *Communications in Statistics*
- Treasurer, Survey Research Methods Section, American Statistical Association

Mary Mulry

- Associate Editor, *Journal of Official Statistics*
- Session Chair, Survey Estimation, 2020 Joint Statistical Meetings

Tapan Nayak

- Associate Editor, *Journal of Statistical Theory and Practice*
- Session Organizer, International Conference on Statistical Distributions and Applications, October 10-12, 2019, Grand Rapids, Michigan
- Session Organizer, IISA 2019 Conference, December 26-30, 2019, Mumbai, India.

Osbert Pang

- Refereed two papers for *Journal of Official Statistics*

Ned Porter

- Session Chair, The 6th IEEE International Conference on Data Science and Advanced Analytics: Research & Application Session on Record Linkage
- Reviewer, IEEE International Conference on Data Science and Advanced Analytics
- Reviewer, 2nd International Workshop on Challenges and Experiences from Data Integration to Knowledge Graphs (DIGK2020) in association with VLDB 2020, the 46th International Conference on Very Large Data Bases

Andrew Raim

- Refereed papers for *Biometrical Journal, Computational Statistics & Data Analysis,* and *Heliyon*.
- Member, Ph.D. Defense Committee, Department of Mathematics and Statistics, University of Maryland, Baltimore County

Kimberly Sellers

- Associate Editor, *The American Statistician*
- Associate Editor, *Journal of Computational and Graphical Statistics*
- Commissioning Editor, *WIREs Computational Statistics*
- Refereed papers for *Journal of Time Series Analysis, Communications in Statistics – Theory and Methods,* and *Computational Statistics and Data Analysis*
- Member, Advisory Board, Summer Program in Research and Learning (SPIRAL), American University

- Member, American Statistical Association (ASA) External Nominations and Awards Committee

Eric Slud

- Associate Editor, *Journal of Survey Statistics and Methodology*
- Associate Editor, *Lifetime Data Analysis*
- Associate Editor, *Statistical Theory and Related Fields*
- Refereed papers for *Computational Statistics and Data Analysis*, *BMC Medical Research Methodology*, *Journal of Royal Statistical Society*, *Series B*, and *Journal of the American Statistical Association*
- Session Organizer, New Approaches to Model-Assisted Survey Estimation with Informative Weights, 2020 Joint Statistical Meetings

Rebecca Steorts

- Associate Editor, *Journal of Survey Statistics and Methodology*
- Associate Editor, *Journal of the American Statistical Association, Applications and Case Studies*
- Associate Editor, *Science Advances*

Thomas Trimbur

- Refereed papers for *Journal of Business Cycle Research* and *Statistical Journal of the International Association for Official Statistics*

Tommy Wright

- Associate Editor, *The American Statistician*
- Member, Board of Trustees, National Institute of Statistical Sciences
- Refereed papers for *Journal of Official Statistics*, *Election Law Journal*, and *Computational Statistics and Data Analysis*
- Member, Task Force on Statistical Significance and Replicability, American Statistical Association
- Member, 2020 Committee of Visitors, Division of Mathematical Sciences, National Science Foundation
- Reviewer, Tenure Review of Faculty Member, Department of Mathematical Sciences, University of Cincinnati

6.3 PERSONNEL NOTES

Brett Moran accepted a position in the Decennial Statistical Studies Division.

Maria Garcia retired following 22 years of Federal Service.

Adam Hall (Ph.D. in statistics, University of Michigan) joined our center as a Census Bureau Postdoctoral Fellow.

APPENDIX A Center for Statistical Research and Methodology FY 2020
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		
6350F01	Address Canvassing In Office		
6450F20	Advertising Campaign		
6550F01	Data Coding/Editing/Imputation		
6550F06	Redistricting Data Program		
6650F01	CCM Planning and Project Management		
6650F20	2020 Evaluations-Planning and Project Management		
6650F25	2030 Planning and Project Management		
6750F01	Administrative Records Data		
	1. 2020 Census Communications Campaign Statistical Analyses-I	Mary Mulry	Nancy Bates
	2. 2020 Census Communications Campaign Statistical Analyses-II	Mary Mulry	Nancy Bates
	3. Supplementing and Supporting Non-Response with Administrative Records	Michael Ikeda	Tom Mule
	4. 2020 Census Privacy Research	James Livsey	Phil Leclerc
	5. Identifying "Good" Administrative Records for 2020 Census NRFU Curtailment Targeting	Darcy Morris	Tom Mule
	6. Experiment for Effectiveness of Bilingual Training	Andrew Raim	Renee Ellis
	7. Unit-Level Modeling of Master Address File Adds and Deletes	Eric Slud	Nancy Johnson
	8. Record-Linkage Support for the Decennial Census	Yves Thibaudeau	David Brown
	9. Coverage Measurement Research	Jerry Maples	Tim Kennel
	10. Assessing Variability of Data Treated by TopDown Algorithm for Redistricting	Tommy Wright	John Abowd
	11. Statistical Modeling to Augment 2020 DAS	Andrew Raim	John Abowd
	American Community Survey (ACS)		
6385F70	*12. Voting Rights Section in 203 Model Evaluation and Enhancements Towards 2021 Determinations*	Carolina Franco	James Whitehorne
	13. Model-based Estimates to Improve Data Confidentiality for ACS Special Tabulations	Jerry Maples	John Abowd
	DEMOGRAPHIC		
TBA	Demographic Statistical Methods Division Special Projects		
	14. Research on Biases in Successive Difference Replication Variance Estimation in Very Small Domains	Eric Slud	Tim Trudell
0906/1444X00	Demographic Surveys Division (DSD) Special Projects		
	15. Data Integration	Edward Porter	Christopher Boniface
TBA	Population Division Projects		
	16. Introductory Sampling Workshop	Tommy Wright	Oliver Fischer
7165019/ 7165020	Social, Economic, and Housing Statistics Division Small Area Estimation Projects		
	17. Research for Small Area Income and Poverty Estimates (SAIPE)	Jerry Maples	Wes Basel
	18. Small Area Health Insurance Estimates (SAHIE)	Ryan Janicki	Wes Basel
	ECONOMIC		
1183X01	Economic Statistical Collection		
	19. Use of Big Data for Retail Sales Estimates	Darcy Morris	Rebecca Hutchinson
	20. Seasonal Adjustment Support	Tucker McElroy	Kathleen McDonald-Johnson
	21. Seasonal Adjustment Software Development and Evaluation	James Livsey	Kathleen McDonald-Johnson
	22. Research on Seasonal Time Series - Modeling & Adjustment Issues	Tucker McElroy	Kathleen McDonald-Johnson
	23. Supporting Documentation & Software for Seasonal Adjustment	James Livsey	Kathleen McDonald-Johnson
	24. Redesign of Economic Sample Surveys (Stratification)	Eric Slud	Justin Z. Smith
	25. Exploring New Seasonal Adjustment & Signal Extraction Methods	James Livsey	Colt Viehdorfer
	26. Classification of Businesses for the NAICS	Emanuel Ben-David	Javier Miranda
	PROGRAM DIVISION OVERHEAD		
0331000	*27. Research Computing*	Chad Russell	Jaya Damineni
	BUREAU OF ECONOMIC ANALYSIS		
TBA	*28. Business Cycle Movements in National Accounts*	Thomas Trimbur	Baoline Chen
	FEDERAL HIGHWAY ADMINISTRATION		
TBA	*29. Modeling & Signal Extraction of Pedestrian & Bicycle Crash Data*	James Livsey	Roya Amjadi

APPENDIX B

FY 2020 PROJECT PERFORMANCE MEASUREMENT QUESTIONNAIRE

CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY

Dear

As a sponsor for the FY2020 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 2020 ***(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 2020 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 FY2020 met expectations.

□ Strongly Agree
□ Agree
□ Disagree
□ Strongly Disagree

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

____

Sponsor Contact Signature Date

Center for Statistical Research and Methodology

Research & Methodology Directorate

STATISTICAL COMPUTING AREA
VACANT

Record Linkage & Machine Learning Research Group
Yves Thibaudeau
Emanuel Ben-David
Xiaoyun Lu
Rebecca Steorts (Duke U.)
Dan Weinberg

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

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

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

MATHEMATICAL STATISTICS AREA
Eric Slud

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

Small Area Estimation Research Group
Jerry Maples
Gauri Datta
Kyle Irimata
Ryan Janicki
Carolina Franco

Time Series & Seasonal Adjustment Research Group
James Livsey
Osbert Pang
Tucker McElroy (Acting)
Soumendra Lahiri (Washington U.)
Anindya Roy (UMBC)
Thomas Trimbur

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

OFFICE OF THE CHIEF
Tommy Wright
Kelly Taylor
Joe Engmark
Adam Hall (P)
Michael Hawkins

(P) Census Bureau Postdoctoral Researcher

January 2021