STATISTICAL EXPERTISE & GENERAL RESEARCH TOPICS
CENTER FOR STATISTICAL RESEARCH & METHODOLOGY
Research & Methodology Directorate
U.S. Bureau of the Census
(FY 2016 – FY 2020)

TOTALLY OR PARTIALLY FUNDED BY
• WORKING CAPITAL FUND
• GENERAL RESEARCH PROJECT

NOVEMBER 2015
To help the Census Bureau continuously improve its processes and data products, general research activity is undertaken in seven broad areas of statistical expertise and general research topics. The activities are supported primarily by the General Research Project and the Working Capital Fund and results from these activities benefit all (decennial, demographic, and economic) programs as well as advance general statistical methodology and practice.

<table>
<thead>
<tr>
<th>Expertise for Collaboration and Research¹</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Missing Data, Edit, and Imputation</td>
<td>1</td>
</tr>
<tr>
<td>2. Record Linkage</td>
<td>3</td>
</tr>
<tr>
<td>3. Small Area Estimation</td>
<td>5</td>
</tr>
<tr>
<td>4. Survey Sampling: Estimation and Modeling</td>
<td>7</td>
</tr>
<tr>
<td>5. Time Series and Seasonal Adjustment</td>
<td>11</td>
</tr>
<tr>
<td>6. Experimentation and Statistical Modeling</td>
<td>14</td>
</tr>
<tr>
<td>7. Simulation and Statistical Modeling</td>
<td>16</td>
</tr>
</tbody>
</table>

¹The Center for Statistical Research & Methodology reviews all research activities and results to ensure that "Census Bureau Statistical Quality Standards" are met and that

- each effort meets a business need of the Census Bureau (motivation, research problem(s), potential applications), which includes how it aligns with the Census Bureau’s strategic plan and the R&M Directorate portfolio management;
- each effort is deeply based in the scientific method and the foundations of statistical science; and
- each effort does not put at risk the Census Bureau’s mission and reputation.
Missing Data, Edit, and Imputation

Motivation: Missing data problems are endemic to the conduct of statistical experiments and data collection projects. The investigators almost never observe all the outcomes they had set out to record. When dealing with sample surveys or censuses, that means individuals or entities omit to respond, or give only part of the information they are being asked to provide. In addition the information provided may be logically inconsistent, which is tantamount to missing. To compute official statistics, agencies need to compensate for missing data. Available techniques for compensation include cell adjustments, imputation and editing, possibly aided by administrative information. All these techniques involve mathematical modeling along with subject matter experience.

Research Problem:

- Compensating for missing data typically involves explicit or implicit modeling. Explicit methods include Bayesian multiple imputation, propensity score matching and direct substitution of information extracted from administrative records. Implicit methods revolve around donor-based techniques such as hot-deck imputation and predictive mean matching. All these techniques are subject to edit rules to ensure the logical consistency of the remedial product. Research on integrating together statistical validity and logical requirements into the process of imputing continues to be challenging. Another important problem is that of correctly quantifying the reliability of predictor in part through imputation, as their variance can be substantially greater than that computed nominally. Specific projects consider (1) nonresponse adjustment and imputation using administrative records, based on propensity and/or multiple imputation models and (2) simultaneous imputation of multiple survey variables to maintain joint properties, related to methods of evaluation of model-based imputation methods.

Potential Applications:

- Research on missing data leads to improved overall data quality and predictors 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 continuously rising cost of conducting censuses and sample surveys, imputation and other missing-data compensation methods aided by administrative records may come to argument actual data collection, in the future.

Accomplishments (October 2014 – September 2015):

- Researched modeling approaches for using administrative records in lieu of Decennial Census field visits in light of forthcoming design decisions.
- Supported the implementation of this research in 2015 Census Test and presented in an invited session at the 2015 Joint Statistical Meetings.
- Investigated the feasibility of using third party (“big”) data from NPD Group to supplement retail sales survey estimates.
- Implemented the model-based procedure Sequential Regression Multivariate Imputation (SRMI) method for missing product data in the Economic Census.
- Initiated and supported the first phase of the integration of Tea to post-process American Community Survey data.

Short-Term Activities (FY 2016):

- Continue researching modeling approaches for using administrative records in lieu of Decennial Census field visits in light of imminent design decisions.
- Investigate the feasibility of using third party (“big”) data from Mastercard to supplement retail sales survey estimates.
- Investigate selective and macro editing large administrative database.
- Complete comparative study of the magnitude of the imputation errors and fraction of missing information when imputing Economic Census product data using ratio imputation and sequential regression multivariate imputation (SRMI).
- Support and expand the functionality of Tea for ACS post-processing activities
- Continue expanding formulae for estimating propensity and efforts in sample surveys from the Mariano-Kadane model or similar parametric models.
Longer-Term Activities (beyond FY 2016):

- Continue researching modeling approaches for using administrative records in lieu of Decennial Census field visits to support future design decisions.
- Research imputation methods for a Decennial Census design that incorporates adaptive design and administrative records to reduce contacts and consequently increases proxy response and nonresponse.
- Research macro and selective editing in the context of large sets of administrative records and high-bandwidth data stream (Big Data).
- Supporting the development and integration of a multifunction data processing system for ACS and other demographic surveys through the open-source software Tea.
- Research model-assisted and model-based estimators based on survey frame information, such as the business register and/or demographic administrative records.
- Research methods for making decisions to mitigate non-response based on interpreting and modeling administrative information.
- Examine the effects of performing large-scale imputation in lieu of completing data collection in the Census Bureau’s largest sample surveys.

Selected Publications:

Contact: Yves Thibaudeau, Maria Garcia, Martin Klein, Darcy Morris, Rolando Rodriguez, Joseph Schafer, Jun Shao, Eric Slud, William Winkler, Xiaoyun Lu.

Funding Sources for FY 2016:
0331 – Working Capital Fund
0925 – General Research Project
Various Decennial, Demographic, and Economic Projects
**Record Linkage**

**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 2006ab, 2008, 2009a, 2013b, 2014a, 2014b; Yancey 2005, 2006, 2007, 2011, 2013) 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 data 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. Specific projects conduct methodological research on multiple-list record linkage, error rates, and statistical inference from linked files.

**Potential Applications:**
- The projects encompass the Demographic, Economic, and Decennial areas and feature linking administrative records with census (decennial and economic) and sample survey data.

**Accomplishments (October 2014 – September 2015):**
- Applied and made updates to record linkage software.
- Presented invited paper “Quality and Analysis of National Files” at the International Symposium on Methodology in Statistics in Ottawa, Ontario Canada.

**Short-Term Activities (FY 2016):**
- Collaboratively define and undertake one major project focused on administrative records and the quality of links.
- Access administrative data for use in record linkage research.
- Train new staff on record linkage methods.
- Teach two-day record linkage short course.

**Longer-Term Activities (beyond FY 2016):**
- Develop methods for adjusting statistical analyses for record linkage error.

**Selected Publications:**


Contact: William E. Winkler, Edward H. Porter, Emanuel Ben-David

Funding Sources for FY 2016: 0331 – Working Capital Fund
0925 – General Research Project
Various Decennial, Demographic, and Economic Projects
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 and for various subpopulations. 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.

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.
- Extend current univariate small-area models to handle multivariate outcomes.
- Development of models to improve uncertainty estimates for design-based estimates near boundaries (e.g., counts near 0, rates near 0 or 1).
- Development of formal methodology for generating small area applications by screening variables from Census Bureau and other federal statistical sample surveys for concordance with American Community Survey variables.

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.
- Extending small area models to estimators of design-base variance.

Accomplishments (October 2014 – September 2015):
- Developed a small area unit level model which uses a large scale sample survey rather than a population roster for prediction.
- Quantified theoretical gains in efficiency in the small area bivariate and times series models over the univariate Fay-Herriot as a function of the noise-to-signal ratio and the correlation over time.
- Created a framework to test the significance of year-to-year changes in small area estimates.
- Developed an EM algorithm to fit Fay-Herriot models with a mixture distribution for the random effect.

Short-Term Activities (FY 2016):
- Investigate alternative methods to estimate MSE for unit level logistic-normal models.
- Develop a principled approach to incorporating higher level design-effects to help improve design-based variances for input into small area models.
- Explore model frameworks to estimate change over time in small area predictions that preserve past small area estimates.
- Test small area models for tract level estimates of poverty rate on artificial simulation samples.
- Set up an automated method and screening rules to find data pairs between two surveys using a bivariate Fay-Herriot model.
- Run simulations to compare single versus double bootstrap procedures in estimating MSE.
- Apply non-lattice spatial models to sample survey data such as the American Community Survey, in order to generate small area estimates and custom region estimates.
Longer-Term Activities (beyond FY 2016):

- Investigate the suitability of area level vs unit level models for small area predictions.
- Modify the tract-level small area models of poverty in school children for school district geographies.
- Examine the distribution of residuals from the Small Area Health Insurance Estimates production model.
- Investigate spatial modeling.

Selected Publications:


Contact: Jerry Maples, Ryan Janicki, Carolina Franco, Gauri Datta, Bill Bell (R&M), Eric Slud, Tucker McElroy

Funding Sources for FY 2016:  
0331 – Working Capital Fund  
0925 – General Research Project  
Various Decennial, Demographic, and Economic Projects
Survey Sampling: Estimation and Modeling

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

Research Problems:
- How to design and analyze sample surveys from "frames" determined by non-probabilistically sampled observational data to achieve representative population coverage?
- What are important areas of research in design-based estimation, model-based estimation, and model-assisted estimation that can benefit Census Bureau data programs?
- How can administrative records, supported by new research on matched survey and administrative lists, be used to increase the efficiency of censuses and sample 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 formulated and solved as optimization problems to avoid the ambiguities resulting from multiple weighting step and to explicitly allow inexact calibration?
- How can we detect and adjust for outliers and influential sample values to improve sample survey estimates?
- What models can aid in assessing the combined effect of all the sources of sampling and nonsampling error, including frame coverage errors and measurement errors, on sample survey estimates?
- What experiments and analyses can inform the development of outreach methods to enhance census response?
- Can unduplication and matching errors be accounted for in modeling frame coverage in censuses and sample surveys?
- What additional information could have been obtained if deleted census persons and housing units had been part of the Census Coverage Measurement (CCM) survey?
- How can small-area or other model-based methods be used to improve interval estimates in sample surveys, to design survey collection methods with lowered costs, or to improve Census Bureau imputation methods?
- Can classical methods in nonparametrics (e.g., using ranks) improve estimates from sample surveys?
- How can we measure and present uncertainty in rankings of units based on sample survey estimates?
- Can BIG DATA improve results from censuses and sample surveys?
- How to develop and use bootstrap methods for expressing uncertainty in estimates from probability sampling.

Potential Applications:
- Improve estimates and reduce costs for household surveys by introducing new design and estimation methods.
- Produce improved ACS small area estimates thorough the use of time series and spatial methods.
- Streamline documentation and make weighting methodology more transparent by applying the same nonresponse and calibration weighting adjustment software across different surveys.
- New procedures for adjusting weights or reported values in the monthly trade surveys and surveys of government employment, based on statistical identification of outliers and influential values, to improve accuracy of estimation monthly level and of month-to-month change.
- Provide a synthesis of the effect of nonsampling errors on estimates of net census coverage error, erroneous enumerations, and omissions and identify the types of nonsampling errors that have the greatest effects. Employ administrative records to improve the estimates of census coverage error.
Use American Community Survey (ACS) data to assess the uncertainty in estimates of foreign-born immigration used by Demographic Analysis (DA) and the Postcensal Estimates Program (PEP) to estimate population size.

Target outreach resources via models to improve the mail response rate in censuses and thereby reduce costs.

Provide a unified computer matching system that can be used with appropriate parameter settings for both the Decennial Census and several Decennial-related evaluations. Reduce decennial census errors by optimally balancing rates of detected and removed census duplicates with rates of coincidental matches.

How to measure and report uncertainty in rankings based on probability samples?

How to develop and use bootstrap methods for expressing uncertainty in estimates from probability samples?

Accomplishments (October 2014 – September 2015):

- Analyzed data from the Commodity Flow Survey to inform decisions by the BTS on conversion to an all- or primarily-electronic data collection strategy, summarizing response rates and data quality by response mode and estimating costs.
- Developed research design for a Pilot Study in American Community Survey concerning curtailment of CAPI follow-up on cases that reach a sufficiently high threshold for a metric of Respondent Burden; began analyzing data results from the Pilot.
- Studied alternative statistical methods for reaching determinations on provision of foreign-language election materials under Section 203 of the Voting Rights Act, and initiating analyses of 2012 ACS data to explore the behavior of those methods.
- Compared the quality of 2010 Census Nonresponse Follow-up proxy responses and administrative records for the same housing units in the same timeframe using the results of the 2010 Census Coverage Measurement Program.
- Synthesized lessons learned from two studies to evaluate survey reports of moves with administrative records to produce a list of challenges researchers must address when using administrative records to evaluate survey responses and identifies steps to take a priori.
- Compared models (primarily logistic regressions) with CUF household size as the dependent variable on the subset of NRFU IDs from Maricopa County where the 2009 IRS 1040 return was present and the UAA flag was blank.
- Investigated potential predictive administrative record covariates for CUF household size based on national data files of NRFU IDs, focusing on potential covariates when 2009 IRS 1040 return data were present and the UAA flag was blank.
- Derived new exact and simple optimal algorithms for allocating the sample considering costs and statement of desired precision.
- Shared methodology for ranking populations based on data from sample surveys with other nations’ statistical agencies.

Short-Term Activities (FY 2016):

- Begin research on the use of “non-probabilistically sampled observational data” complemented by results from probability sample surveys.
- Code general-purpose software for weight adjustment for nonresponse and frame errors via calibration with some inexact constraints using quadratic programming to enforce bounds on multiplicative weight changes.
- Compare the relative accuracy of occupancy status and responses for housing units enumerated in the 2015 Census Nonresponse Follow-up versus the administrative records, using data from the 2010 Census Coverage Measurement Program.
- Analyze the effects of outreach on response to censuses and sample surveys.
- Explore the possibility of replacing the confidence intervals used in ACS for another direct method.
- Further investigate the potential benefit of using small area time series estimation models on ACS data.
- Continue to study properties of the exact optimal allocation method and its application.
- Continue to investigate ranking methods based on sample survey estimates.
- Begin research into the benefits of the bootstrap in results from sample surveys.
- Implement some ranking methodology and visualizations on Census Bureau estimates.

Longer-Term Activities (beyond FY 2016):

- Investigate the properties and applicability of alternative design and estimation methods for household sample surveys: both design-based and model-based.
- Develop sample survey methodology to augment the expanded use of nonprobability, administrative and observational data for the estimation of household and business populations.
- Investigate the use of design effect adjustments in confidence intervals for estimates based on complex survey data.
- Expand the search for methods appropriate for estimation methodology in the presence of influential values in sample surveys.
• Develop methodology for estimation and for assessing the quality of estimates using multiple sources of data, such as multiple sample surveys, administrative records systems, and/or a census.
• Develop methods for estimating the uncertainty due to nonsampling errors in estimates from censuses, sample surveys, and administrative records.
• Continue and expand alternative methods of single-stage survey weighting based on propensity-adjusted estimating equations and optimization methods to accommodate both exact and approximate calibration constraints.
• Investigate the relative accuracy of occupancy status and responses for housing units enumerated in the 2015 Census Test Nonresponse Followup with the accuracy of the administrative records using data collected in the 2015 Evaluation Followup.
• Develop statistical models for using administrative records (AR) in the 2020 Census NRFU process. Help use these models and other information to determine the most effective use of AR in the NRFU to reduce costs while maintaining quality.
• Expand the applicability of a realistic simulation utility for ACS, based on a frame constructed from census and ACS data, for the purpose of ACS estimation methods development, including methods for reporting confidence intervals and for small area estimation.
• Investigate ranking methods based on sample survey estimates as well as the use of ranks in sample surveys.

Selected Publications:
Period Estimates.” Published online, *Journal of the International Association of Official Statistics*.

Contact: Eric Slud, Robert Ashmead, Mary Mulry, Michael Ikeda, Patrick Joyce, Martin Klein, Ned Porter, Tommy Wright

**Funding Sources for FY 2016:**

0331 – Working Capital Fund
0925 – General Research Project
Various Decennial, Demographic, and Economic Projects
Time Series and Seasonal Adjustment

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

**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.
- Need better diagnostics and measures of estimation and adjustment quality, e.g., for model-based seasonal adjustment.
- Develop a viable multivariate seasonal adjustment methodology that can handle a large number of series, such that the aggregation constraints of multiple time series can be directly accounted for.
- Develop and extend seasonal adjustment methods to time series observed at higher frequencies, e.g., weekly data.

**Potential Applications**
- Potential applications encompass the Decennial, Demographic, and Economic areas.

**Accomplishments (October 2014 – September 2015):**
- Completed new methodology of extreme-value adjustment for seasonal adjustment, with application to New Zealand agricultural data.
- Developed an R user-interface for X-13ARIMA-SEATS, allowing for greater usability and communicability of seasonal adjustments.
- Developed new estimation methods for vector time series models, allowing for parameter constraints.
- Developed a hybrid model-based seasonal adjustment method using SEATS in a moving window framework, implemented in R.
- Determined that the model-based trading day F-statistic has better power and a lower false detection rate than other available trading day diagnostics.

**Short-Term Activities (FY 2016):**
- Continue development of X-13ARIMA-SEATS needed to publish SEATS seasonal adjustments of selected Census Bureau series, and re-engineer the software around a more modern computing paradigm.
- Continue investigation of more exotic seasonal models, including seasonally heteroscedastic and season-specific models, and develop the corresponding seasonal adjustment methods.
- Expand research on multivariate seasonal adjustment, seasonal long memory modeling of time series, and spectral peak detection.
- Improve the speed and stability of likelihood optimization in X-13ARIMA-SEATS.
- Continue examining methods for estimating trading day regressors with time-varying coefficients, and determine which Census Bureau series are amenable to moving trading day adjustment.

**Longer-Term Activities (beyond FY 2016):**
- Develop models that allow for heavy-tailed marginal distributions, to allow for extreme values in seasonal adjustment algorithms.
- Study the modeling of seasonal time series measured at higher sampling frequencies, such as weekly and daily frequencies.
- Develop a fundamental inquiry into the nature of seasonality, which goes beyond parametric or diagnostic paradigms, and apply the definition to addressing the outstanding issues of seasonal adjustment adequacy, sufficiency, and comparability.
- Develop a comprehensive framework for comparing and evaluating modeling paradigms for time series, taking into account non-stationary structures, non-Gaussian structures, and non-parametric structures, transcending the narrowly residual-based
diagnostic philosophy.
• Examine the feasibility of using the Bayes seasonal adjustment method on Census Bureau time series.

Selected Publications:


**Contact:** Tucker McElroy, Brian C. Monsell, James Livsey, Osbert Pang, Anindya Roy, Bill Bell (R&M), Thomas Trimbur.

**Funding Sources for FY 2016:**
- 0331 – Working Capital Fund
- 0925 – General Research Project
- Economic Projects
**Experimentation and Statistical 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 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. Uses simulation studies.
- Research methods to provide principled measures of statistical variability for constructs like the POP Division’s Population Estimates.
- 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.

**Accomplishments (October 2014 – September 2015):**
- Developed code to carry out exhaustive variable selection for the MAF Error Model. Ran selection process on Decennial Statistical Studies Division (DSSD) address canvassing database plus six supplemental data sources. Considered two-way interactions for all main effects. Used randomized quantile residuals and other evaluations to study the selected zero-inflated negative binomial model.
- Started work on spatial modeling with DSSD staff for the MAF Error Model. Investigated Bayesian conditional autoregressive (CAR) models for counts in small geographies (county level).
- Delivered models to DSSD for prediction of 2015 Address Validation Test (AVT) field test.
- Assisted with handoff of the Census Enterprise Model from MITRE to the Census Bureau, and helped to develop programs to simulate address canvassing within the microsimulator.
- Presented a series of lectures on survey sampling and experimental design for a general Census Bureau audience.

**Short-Term Activities (FY 2016):**
- Begin investigation of methodology to embed experiments into sample surveys.
- Begin to investigate the use of the bootstrap in sample surveys.

**Longer-Term Activities (beyond FY 2016):**
- Continue investigation of methodology to embed experiments into sample surveys.
- Use zero-inflated model results to help inform data collection efforts for expanding the universe file used by both the MAF Error Model (MEM) team and the GSSI Targeted Address Canvassing Research, Modeling, and Classification (TRMAC) team.
• Apply existing or develop new methodologies for properly incorporating spatial components in the ongoing frame improvement research.

**Selected Publications:**

**Contact:** Andrew Raim (x37894), Thomas Mathew, Kimberly Sellers.

**Funding Sources for FY 2016:**

0331 – Working Capital Fund
0925 – General Research Project
Various Decennial and Demographic Projects
Simulation and Statistical Modeling

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 computationally intensive statistical methods. Modeling also enables one to efficiently use all available information when producing estimates. Such studies can benefit from software such as Tea for data processing. Statistical disclosure avoidance methods are also developed and properties studied.

Research Problems:
- Systematically develop an environment for simulating complex surveys that can be used as a test-bed for new data analysis methods.
- Develop flexible model-based estimation methods for 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.
- Continue to formalize the codebase and user interfacing for Tea, especially within the context of the current enterprise environment.
- 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.
- Tea provides modeling and editing flexibility, especially with a focus on incorporating administrative data.
- Using an environment for simulating complex 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.

Accomplishments (October 2014 – September 2015):
- Developed new model-based methodology for analyzing singly imputed synthetic data under exponential and univariate normal models. Developed similar methods for multivariate normal data assuming all variables are synthesized, and multiple linear regression models when the response variable is synthesized.
- Developed new methodology for using noise multiplication as an alternative to top coding for statistical disclosure control under a log-normal model where the log-scale mean is modeled as a linear regression.
- Created an initial artificial population for simulating the Monthly Wholesale Trade Survey over a two-year period and used it to evaluate imputation and estimation methodology.
- Used discrete time logistic regression to develop daily response propensity models for the National Crime Victimization Survey.
- Applied bootstrap methods to quantify uncertainty in rankings of states based on American Community Survey data along with visualizations.

**Short-Term Activities (FY 2016):**
- Develop exact model-based methods for analyzing singly imputed synthetic data under linear regression models where both response and regressor variables are synthesized.
- Evaluate properties of singly imputed synthetic data when the analysis model differs from the imputation model.
- Develop methodology for using noise infusion for statistical disclosure control for data containing zeros.
- Refine the artificial population created for simulating the Monthly Wholesale Trade Survey.
- Add additional modeling and editing flexibility to Tea.
- Refine response propensity models for the National Crime Victimization Survey.
- Evaluate properties of bootstrap methods for sample survey data.

**Longer-Term Activities (beyond FY 2016):**
- Develop noise infusions methods for statistical disclosure control under nonparametric models, and develop likelihood-based methods for analyzing singly imputed synthetic data when values are synthesized under a non-ignorable data masking mechanism.
- Develop bootstrap methods for analyzing synthetic and noise infused data.
- Study ways of quantifying the privacy protection/data utility tradeoff in statistical disclosure control.
- Develop methods for using bootstrap calibration for estimation in sample survey data.
- Create an environment for simulating complex aspects of economic/demographic surveys.
- Study properties of bootstrap methodology for quantifying uncertainty in statistical rankings and refine visualizations.
- Improve and seek new applications for Tea.

**Selected Publications:**

**Contact:** Martin Klein, Isaac Dompreh, Brett Moran, Rolando Rodriguez, Nathan Yau.

**Funding Sources for FY 2016:**
- 0331 – Working Capital Fund
- 0925 – General Research Project
- Various Decennial, Demographic, and Economic Projects
<table>
<thead>
<tr>
<th>96 10 10</th>
<th>STATISTICAL COMPUTING AREA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Winkler (Acting)</td>
<td>5K005...x34729</td>
</tr>
<tr>
<td>Alisha Armas</td>
<td>5K110C...x36637</td>
</tr>
</tbody>
</table>

**Machine Learning & Computational Statistics Research Group**

| Bill Winkler | 5K005...x34729 |
| Emanuel Ben-David | 5K102F...x37275 |
| Xiaoyun Lu | 5K104B...x35395 |

**Statistical Computing Applications & Data Visualization Research Group**

| Martin Klein (Acting) | 5K106F...x37856 |
| Rolando Rodriguez | 5K101A...x31816 |
| Nathan Yau (FLOWINGDATA.COM) | 5K110E...x37664 |

**Missing Data Methods Research Group**

| Yves Thibaudeau | 5K105...x31706 |
| Douglas Galagate (S) | 5K112C...x34985 |
| Maria Garcia | 5K102A...x31703 |
| Darcy Morris | 5K101B...x33989 |
| Jun Shao (U. of WI) | 5K413...x33101 |
| VACANT | |

**Research Computing Systems & Applications Group**

| Chad Russell | 5K006...x33215 |
| Tom Petkunas | 5K102C...x33216 |
| Ned Porter | 5K104A...x31798 |

**Simulation & Modeling Research Group**

| Martin Klein | 5K106F...x37856 |
| Isaac Dompreh | 5K106C...x36801 |
| Brett Moran | 5K106E...x39986 |
| Derrick Simmons (S) | 5K507G...x36231 |

<table>
<thead>
<tr>
<th>96 20 20</th>
<th>MATHEMATICAL STATISTICS AREA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eric Sud</td>
<td>5K004...x34991</td>
</tr>
<tr>
<td>Erica Magruder (HRD)</td>
<td>5K110A...x36751</td>
</tr>
</tbody>
</table>

**Sampling & Estimation Research Group**

| Eric Sud (Acting) | 5K004...x34991 |
| Robert Ashmead | 5K112A...x31564 |
| Mike Ikeda | 5K106A...x31756 |
| Patrick Joyce | 5K106B...x36793 |
| Mary Mulry | 5K415...x31759 |

**Small Area Estimation Research Group**

| Jerry Maples | 5K003...x32873 |
| Gauri Datta (U. of Georgia) | 5K112B...x33426 |
| Carolina Franco | 5K112F...x39959 |
| Ryan Janicki | 5K102B...x35725 |

**Time Series Research Group**

| Brian Monsell | 5K002...x31721 |
| Osbert Pang | 5K102D...x30252 |
| VACANT | |
| Tucker McElroy | 5K103...x33227 |
| James Livsey | 5K104F...x33517 |
| Anindya Roy (UMBC) | 5K104D...x33591 |
| Thomas Trimbur | 5K104E...x36864 |
| Service Number | 5K018...x31649 |

**Experimenation & Modeling Research Group**

| Tommy Wright (Acting) | 5K108...x31702 |
| Thomas Mathew (UMBC) | 5K112E...x35537 |
| Andrew Raim | 5K112D...x37894 |
| Kimberly Sellers (Georgetown U.) | 5K102E...x39808 |

<table>
<thead>
<tr>
<th>96 00 00</th>
<th>OFFICE OF THE CHIEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tommy Wright, Chief</td>
<td>5K108...x31702</td>
</tr>
<tr>
<td>Kelly Taylor</td>
<td>5K108C...x34896</td>
</tr>
<tr>
<td>Lauren Emanuel</td>
<td>5K110D...x39862</td>
</tr>
<tr>
<td>Michael Hawkins</td>
<td>5K110B...x31691</td>
</tr>
<tr>
<td>Michael Leibert</td>
<td>5K106D...x32393</td>
</tr>
<tr>
<td>Andrew Perry (S)</td>
<td>5K104C...x31219</td>
</tr>
<tr>
<td>(S) Student</td>
<td></td>
</tr>
</tbody>
</table>