



**STATISTICAL EXPERTISE &
GENERAL RESEARCH TOPICS**
CENTER FOR STATISTICAL RESEARCH & METHODOLOGY
Research & Methodology Directorate
U.S. Bureau of the Census
(FY 2018 – FY 2022)

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

SEPTEMBER 2017



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 of 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.

Expertise for Collaboration and Research¹		Page
1.	Missing Data, Edit, and Imputation	1
2.	Record Linkage	3
3.	Small Area Estimation	5
4.	Survey Sampling: Estimation and Modeling	7
5.	Time Series and Seasonal Adjustment	11
6.	Experimentation and Statistical Modeling	14
7.	Simulation and Statistical Modeling	17

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 2016 – September 2017):

- Showed how to use log-linear models to improve the efficiency (reducing the sampling error) of longitudinal estimators of quarterly changes in labor force status and healthcare coverage. Showed how these estimators can be implemented for labor force and healthcare coverage measurement using data from the Survey of Income and Program Participation (*Annals of Applied Statistics*).
- Researched modeling approaches for using administrative records in lieu of Decennial Census field visits due to forthcoming design decisions (*Public Opinion Quarterly* and *Statistical Journal of the IAOS*).
- Investigated the feasibility of using third party (“big”) data from First Data –a large payment processor to supplement and/or enhance retail sales estimates in the Monthly/Annual Retail Trade Survey (MRTS and ARTS).
- Researched, developed, and evaluated methods for raking balance complexes in the Quarterly Financial Report (QFR) when items are negative or there is subtraction in the balance complexes.
- Collaborated in the development of four separate alternative methods to raking balance complexes in the Standard Economic Processing System (StEPS) when detail items are negative or there is subtraction in the balance complexes.
- Set-up the problem of augmenting the exports and patents datasets with variables from the Business Register (BR) as a missing data problem and proposed two separate approaches: Statistical Matching and the multiple imputation procedure Sequential Regression Multivariate Imputation (SRMI).
- Developed a system that generates essentially new implied edits based on given explicit edits.

Short-Term Activities (FY 2018):

- Extend the analysis and estimation of changes in the labor force status using log-linear models to the Current Population Survey.
- Continue researching modeling approaches for using administrative records in lieu of Decennial Census field visits due to imminent design decisions.
- Continue to investigate the feasibility of using third party (“big”) data from various available sources to supplement and/or

enhance retail sales estimates in the Monthly/Annual Retail Trade Survey (MRTS and ARTS).

- Continue research on augmenting export transactions and patents data files by adding variables from the business register (BR).
- Continue research on alternative methods for raking balance complexes when variables are allowed to be negative or there is subtraction in the balance complex.
- Collaborate with Economic Directorate in setting-up parameters, adjusting reliability weights, and transitioning to implementation of alternative raking methods for QFR processing in the Standard Economic Processing System (StEPS)
- Continue research on applying Bayesian editing methods developed by Hang Kim et al. (2015) to developing synthetic economic census data.
- Continue work on heuristic methods for edit generations.

Longer-Term Activities (beyond FY 2018):

- Extend small area estimation modeling for longitudinal data (survey and/or third party) in presence of attrition and/or other type of missing data using log-linear models.
- Extend the modeling of propensity jointly with the accuracy of administrative sources.
- Continue researching modeling approaches for using administrative records in lieu of Decennial Census field visits to support future design decisions.
- Research practical ways to apply decision theoretic concepts to the use of administrative records (versus personal contact or proxy response) in the Decennial Census.
- Research joint models for longitudinal count data and missing data (e.g. drop out) using shared random effects to measure the association between propensity for nonresponse and the count outcome of interest.
- 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).
- Continue collaboration on researching methods for data integration of the exports and patents data files with the Business Register (BR).
- Evaluate the results of data corrections in the Standard Economic Processing System (StEPS) using new raking algorithms for adjusting balance complexes.
- Continue research on edit procedures.
- Investigate why some of the newly developed alternative methods for raking lead to lower weighted totals than the existing StEPS raking method, apply the methodology to additional balance complexes from the QFR, and research the application to balance complexes from other Economic Census surveys.

Selected Publications:

- Bechtel, L., Morris, D.S., and Thompson, K.J. (2015). "Using Classification Trees to Recommend Hot Deck Imputation Methods: A Case Study." In *FCSM Proceedings*. Washington, DC: Federal Committee on Statistical Methodology.
- Garcia, M., Morris, D.S., and Diamond, L.K. (2015). "Implementation of Ratio Imputation and Sequential Regression Multivariate Imputation on Economic Census Products." *Proceedings of the Joint Statistical Meetings*.
- Morris, D.S., Keller, A., and Clark, B. (2016). "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census." *Statistical Journal of the International Association for Official Statistics*, 32(2): 177-188.
- Morris, D. S. (2017). "A Modeling Approach for Administrative Record Enumeration in the Decennial Census," *Public Opinion Quarterly: Special Issue on Survey Research, Today and Tomorrow*, 81(S1): 357-384.
- Thibaudeau Y., Slud, E., and Gottschalck, A. O. (2017). "Modeling Log-Linear Conditional Probabilities for Estimation in Surveys," *Annals of Applied Statistics* 11(2), 680-697.
- Thibaudeau, Y. (2002). "Model Explicit Item Imputation for Demographic Categories," *Survey Methodology*, 28(2), 135-143.
- Winkler, W. E. (2008). "General Methods and Algorithms for Imputing Discrete Data under a Variety of Constraints," *Research Report Series (Statistics #2008-08)*, Statistical Research Division, U.S. Census Bureau, Washington DC.
- Winkler, W. and Garcia, M. (2009). "Determining a Set of Edits," *Research Report Series (Statistics #2009-05)*, Statistical Research Division, U.S. Census Bureau, Washington, DC.

Contact: Yves Thibaudeau, Maria Garcia, Martin Klein, Darcy Morris, Jun Shao, Eric Slud, William Winkler, Xiaoyun Lu

Funding Sources for FY 2018: 0331 – Working Capital Fund / 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 2006a,b, 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 2016 – September 2017):

- Applied and made updates to record linkage software. Software continues to be recognized for its accuracy and speed.
- Located and slightly updated thirty programs of initial methods for record linkage error-rate estimation.
- Gave a principal technical talk “Computational Methods for Cleaning and Analyzing Administrative Files” at the Isaac Newton Institute.
- Wrote the chapter “Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Editing/Imputation” that will appear in a Wiley monograph on administrative records edited by Chun and Larsen.

Short-Term Activities (FY 2018):

- Train people to update and maintain the 45+ programs for record linkage and related data preparation.
- Lead a task force that will decide the main record linkage research problems that the Census Bureau will address.
- Conduct research and share findings on the effects of linkage error on statistical analyses.
- Conduct research on record linkage error-rate estimation, particularly for unsupervised and semi-supervised situations.

Longer-Term Activities (beyond FY 2018):

- Develop methods for adjusting statistical analyses for record linkage error. We believe that seventeen papers in fifty-two years provide between five and twenty percent of the solution.

Selected Publications:

- Alvarez, M., Jonas, J., Winkler, W. E., and Wright, R. “Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project,” *Electronic Voting Technology*.
- Herzog, T. N., Scheuren, F., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*, New York, NY: Springer.
- Herzog, T. N., Scheuren, F., and Winkler, W. E. (2010). “Record Linkage,” in (Y. H. Said, D. W. Scott, and E. Wegman, eds.) *Wiley Interdisciplinary Reviews: Computational Statistics*.
- Winkler, W. E. (2006a). “Overview of Record Linkage and Current Research Directions,” *Research Report Series (Statistics #2006-02)*, Statistical Research Division, U.S. Census Bureau, Washington, DC.

- Winkler, W. E. (2006b). "Automatically Estimating Record Linkage False-Match Rates without Training Data," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, CD-ROM.
- Winkler, W. E. (2008). "Data Quality in Data Warehouses," in (J. Wang, Ed.) *Encyclopedia of Data Warehousing and Data Mining (2nd Edition)*.
- Winkler, W. E. (2009a). "Record Linkage," in (D. Pfeffermann and C. R. Rao, eds.) *Sample Surveys: Theory, Methods and Inference*, New York: North-Holland, 351-380.
- Winkler, W. E. (2009b). "Should Social Security numbers be replaced by modern, more secure identifiers?," *Proceedings of the National Academy of Sciences*.
- Winkler, W. E. (2010). "General Discrete-data Modeling Methods for Creating Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties," <http://www.census.gov/srd/papers/pdf/rrs2010-02.pdf>.
- Winkler, W. E. (2011). "Machine Learning and Record Linkage" in *Proceedings of the 2011 International Statistical Institute*.
- Winkler, W. E. (2013). "Record Linkage," in *Encyclopedia of Environmetrics*. J. Wiley.
- Winkler, W. E. (2013). "Cleanup and Analysis of Sets of National Files," Federal Committee on Statistical Methodology, Proceedings of the Bi-Annual Research Conference, http://www.copafs.org/UserFiles/file/fcsm/J1_Winkler_2013FCSM.pdf, https://fcsm.sites.usa.gov/files/2014/05/J1_Winkler_2013FCSM.pdf
- Winkler, W. E. (2014a). "Matching and Record Linkage," *Wiley Interdisciplinary Reviews: Computational Statistics*, <http://wires.wiley.com/WileyCDA/WiresArticle/wisId-WICS1317.html>, DOI: 10.1002/wics.1317, available from author by request for academic purposes.
- Winkler, W. E. (2014b). "Very Fast Methods of Cleanup and Statistical Analysis of National Files," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM.
- Winkler, W. E. (2015). "Probabilistic Linkage," in (H. Goldstein, K. Harron, C. Dibben, eds.) *Methodological Developments in Data Linkage*, J. Wiley: New York.
- Winkler, W. E. (2018 to appear). "Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Editing/Imputation," in (A.Y. Chun and M. D. Larsen, eds.) *Administrative Records for Survey Methodology*, J. Wiley, New York: NY.
- Winkler, W. E., Yancey, W. E., and Porter, E. H. (2010). "Fast Record Linkage of Very Large Files in Support of Decennial and Administrative Records Projects," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Alexandria, VA.
- Yancey, W. E. (2005). "Evaluating String Comparator Performance for Record Linkage," *Research Report Series (Statistics #2005-05)*, Statistical Research Division, U.S. Census Bureau, Washington, DC.
- Yancey, W. E. (2007). "BigMatch: A Program for Extracting Probable Matches from a Large File," *Research Report Series (Computing #2007-01)*, Statistical Research Division, U.S. Census Bureau, Washington, DC.

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

Funding Sources for FY 2018: 0331 – Working Capital Fund / 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.
- Using the American Community Survey to improve the precision of estimates from other smaller surveys.

Accomplishments (October 2016 – September 2017):

- Developed a unit-level small area projection model to estimate state level disability rates from the Survey of Income and Program Participation augmented by the American Community Survey.
- Developed a small area model for tracts using a generalized Poisson distribution.
- Via simulation, application, and theory, studied the properties of functional and structural measurement models in small area estimation and the impact of ignoring or incorrectly specifying measurement error.
- Proposed a way to improve the coverage of simple direct methods of confidence interval construction for proportions in complex surveys by improving estimation of the effective sample size using superpopulation model assumptions. Used this approach to improve the coverage of alternatives to Wald-type intervals, the most widely used interval methods in applications, and showed via a comprehensive simulation with a factorial design that many of the alternatives perform much better in term of achieving the nominal coverage.
- Developed an area-level model for rates using Beta regression.

Short-Term Activities (FY 2018):

- Rebuild the design-based samples from the Artificial Population dataset (ACS 2008-2012) to allow for both area- and unit-level models to be evaluated.
- Compare different generalized Poisson sampling distribution models for tract level estimates of child poverty.
- Extend the Beta small area model for rates to an Inflated Zero-One Beta model and apply model to health insurance estimates for SAHIE.
- Evaluate different small area models for rates using the artificial population samples.

Longer-Term Activities (beyond FY 2018):

- Investigate unit-level small area models for multivariate outcomes.
- Develop model-based small area estimates of poverty in school-aged children for school districts.
- Incorporate spatial modeling into the small area effects and develop tests for their necessity.
- Develop methods for estimating the change over time of small area parameters and for constructing interval estimates for this change.

Selected Publications:

- Arima, S., Bell, W. R., Datta, G. S., Franco, C. and Liseo, B. (In Press). "Multivariate Fay-Herriot Bayesian estimation of Small Area Means Under Functional Measurement Error," *Journal of the Royal Statistical Society--Series A*.
- Franco, C. and Bell, W. R. (2013). "Applying Bivariate/Logit Normal Models to Small Area Estimation," In *JSM Proceedings, Survey Research Methods Section*. Alexandria, VA: American Statistical Association. 690-702.
- Franco, C. and Bell, W. R. (2015). "Borrowing information over time in binomial/logit normal models for small area estimation," *Joint issue of Statistics in Transition and Survey Methodology*. 16, 4, 563-584.
- Datta, G., Ghosh, M., Steorts, R. and Maples, J. (2011). "Bayesian Benchmarking with Applications to Small Area Estimation," *TEST, Volume 20, Number 3*, 574-88.
- Huang, E., Malec, D., Maples J., and Weidman, L. (2007). "American Community Survey (ACS) Variance Reduction of Small Areas via Coverage Adjustment Using an Administrative Records Match," *Proceedings of the 2006 Joint Statistical Meetings*, American Statistical Association, Alexandria, VA, 3150-3152.
- Janicki, R. (2011). "Selection of prior distributions for multivariate small area models with application to small area health insurance estimates." *JSM Proceedings, Government Statistics Section*. American Statistical Association, Alexandria, VA.
- Janicki, R. and Vesper, A. (2017). "Benchmarking techniques for reconciling small area models at distinct geographic levels." *Statist. Methods Appl*, DOI: <https://doi.org/10.1007/s10260-017-0379-x>
- Janicki, R (2016). "Estimation of the difference of small area parameters from different time periods". *Research Report Series (Statistics #2016-01)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC.
- Joyce, P. and Malec, D. (2009). "Population Estimation Using Tract Level Geography and Spatial Information," *Research Report Series (Statistics #2009-3)*, Statistical Research Division, U.S. Census Bureau, Washington, DC.
- Malec, D. (2005). "Small Area Estimation from the American Community Survey using a Hierarchical Logistic Model of Persons and Housing Units," *Journal of Official Statistics*, 21 (3), 411-432.
- Malec, D. and Maples, J. (2008). "Small Area Random Effects Models for Capture/Recapture Methods with Applications to Estimating Coverage Error in the U.S. Decennial Census," *Statistics in Medicine*, 27, 4038-4056.
- Malec, D. and Müller, P. (2008). "A Bayesian Semi-Parametric Model for Small Area Estimation," in *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* (eds. S. Ghoshal and B. Clarke), Institute of Mathematical Statistics, 223-236.
- Maples, J. and Bell, W. (2007). "Small Area Estimation of School District Child Population and Poverty: Studying Use of IRS Income Tax Data," *Research Report Series (Statistics #2007-11)*, Statistical Research Division, U.S. Census Bureau, Washington, DC.
- Maples, J. (2011). "Using Small-Area Models to Improve the Design-Based Estimates of Variance for County Level Poverty Rate Estimates in the American Community Survey," *Research Report Series (Statistics #2011-02)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC.
- Maples, J. (In Press). "Improving Small Area Estimates of Disability: Combining the American Community Survey with the Survey of Income and Program Participation," *Journal of the Royal Statistical Society—Series A*.
- Slud, E. and Maiti, T. (2006). "Mean-Squared Error Estimation in Transformed Fay-Herriot Models," *Journal of the Royal Statistical Society Series B*, 239-257.
- Slud, E. and Maiti, T. (2011). "Small-Area Estimation Based on Survey Data from Left-Censored Fay-Herriot Model," *Journal of Statistical Planning & Inference*, 3520-3535.

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

Funding Sources for FY 2018: 0331 – Working Capital Fund / 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. To make census data products based jointly on administrative and survey data fully representative of the general population, as our current surveys are, new sampling designs and analysis methods will have to be developed.
- How can administrative records, supported by new research on matched survey and administrative lists, be used to increase the efficiency of censuses and sample surveys?
- How can inclusion in observational or administrative lists be modeled jointly with indicator and mode of survey response, so that traditional survey methods can be extended to merged survey and non-survey data?
- Can non-traditional design methods such as adaptive sampling be used to improve estimation for rare characteristics and populations?
- How can time series and spatial methods be used to improve ACS estimates or explain patterns in the data?
- Can generalized weighting methods be formulated and solved as optimization problems to avoid the ambiguities resulting from multiple weighting step and to explicitly allow inexact calibration?
- 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?
- 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 through the use of time series and spatial methods, where those methods improve upon small area methods using covariates recoded from temporal and spatial information.
- 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.

- Measure and report uncertainty in rankings in household and economic sample surveys.
- Develop bootstrap methods for expressing uncertainty as an alternative source of published variance estimates and as a check on existing methods of producing variances in Census Bureau sample surveys.

Accomplishments (October 2016 – September 2017):

- Developed model diagnostics and variance estimation methodology for small area prediction within language minority groups in connection with the determinations of alternative language election assistance by jurisdiction and American Indian Area under *Section 203 of the Voting Rights Act*. The variance methodology has potential for application in many different sample survey settings that combine design-based and model-based estimates
- Analyzed the potential of market segmentation from an external source to provide useful information to the 2020 Census communications campaign.
- Developed alternative methods of ranking census tracts for their potential to be influenced by decennial census outreach operations.
- Developed a conceptual framework for designing, testing, and evaluating policies for curtailment of followup in household surveys, modelled on research done with and by American Community Survey staff on CATI and CAPI contact history paradata.
- Developed methodology for the use of Administrative Records to supplement and replace decennial census data collection, using models and data substitution.
- Developed methodology for comparing the quality of results of different options for census-taking for use in a cost-benefit analysis.
- Contributed methodology for assessment of side-by-side test of influential value detection for an economic sample survey.
- Applied an external database not commonly used in survey research in analyses that produced results strongly suggesting: (1) the Census Bureau's Low Response Score is a good metric to predict Census self-response at the tract level, even with the addition of the Internet as a response mode, and (2) an external segmentation designed for commercial marketing could be useful in the planning the 2020 Census communications campaign.
- Developed and demonstrated that exact optimal sample allocation algorithms can produce savings in the redesign of the Census Bureau's Service Annual Survey which takes place following each Economic Census.
- Completed theory and visualizations in comparing several populations with overlapping and non-overlapping confidence intervals. Produced and published associated software on CRAN.
- Conducted two Sampling Workshops: (1) a two-week session at Census Bureau Headquarters for six participants from statistical agencies in Mozambique, Egypt, and Pakistan (USAID funding) and (2) a one-week session in Kigali, Rwanda for sixteen statisticians at the National Institute of Statistics of Rwanda (formed in 2005) which provided funding for participants.

Short-Term Activities (FY 2018):

- Extend methodology for partially model-based estimation methods in longitudinal surveys to include additional time periods, a geographic small-area component, and a rotating panel design as in the Current Population Survey.
- Develop differentially private methods for imputing geography down to the block level given a differentially private national level synthetic histogram, for application in Census 2020. Investigate methods to create differentially private synthetic households from person-level data.
- Continue to investigate ranking methods based on sample survey estimates.
- Continue research into the benefits of the bootstrap compared with other variance estimation methods in sample surveys.
- Continue research into the combination of census data collection with administrative records, and evaluation of coverage based on these methods.
- Continue research into diagnostics and goodness of fit methodology for generalized linear mixed models, with particular attention to applications in response propensity models and in small area estimation based on sample surveys.
- Analyze the potential for a market segmentation from an external source to aid in providing useful information about problems in the Census enumeration of young children.
- Contribute to statistical analyses that support the 2020 Census communications campaign.
- Contribute to the assessment of a side-by-side test of influential value detection methodology for an economic sample survey.

Longer-Term Activities (beyond FY 2018):

- To create a disclosure avoidance system for releasing detailed summary tables or microdata from a large-scale survey or census under the differential privacy paradigm with an adjustable privacy budget.
- Continue development of methodology for survey estimation for Census surveys administered under protocols guaranteeing differential privacy.
- Develop methods for survey data-collection and analysis that combine probability sampling with observational administrative and commercial databases, in such a way as to provide coverage and representativeness guarantees.
- Develop improved methodology for sample survey inference based on survey data linked imperfectly to observational/administrative/commercial data, with attention to models for linkage error-rate estimation.
- Develop methodology to increase understanding of the undercoverage of young children in censuses and surveys and contribute to improving coverage of this group.
- Investigate ranking methods and visualizations based on sample survey estimates as well as the use of ranks in sample surveys.
- Develop methodology to increase understanding of the undercoverage of young children in censuses and surveys and contribute to improving coverage of this group.

Selected Publications:

- Ashmead, R., Slud, E., and Hughes, T. (In Press), “Adaptive Intervention Methodology for Reduction of Respondent Contact Burden in the American Community Survey,” *Journal of Official Statistics*.
- Ashmead, R. and Slud, E. (2017), “Small area model diagnostics and validation with applications to the Voting Rights Act Section 203”, *Proceedings of Survey Research Methods Section*, American Statistical Association, Alexandria, VA.
- Franco, C., Little, R., Louis, T., and Slud, E. (2014). “Coverage Properties of Confidence Intervals for Proportions in Complex Sample Surveys,” *Proceedings of Survey Research Methods Section*, American Statistical Association, Alexandria, VA.
- Griffin, D., Slud, E., and Erdman, C. (2014). “Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation - Phase 3 Results,” *ACS Research and Evaluation Memorandum #ACS 14-RER-28*.
- Hogan, H. and Mulry, M. H. (2014). “Assessing Accuracy of Postcensal Estimates: Statistical Properties of Different Measures,” in N. Hogue (Ed.), *Emerging Techniques in Applied Demography*. Springer. New York.
- Hunley, Pat. (2014). “Proof of Equivalence of Webster’s Method and Willcox’s Method of Major Fractions,” *Research Report Series (Statistics #2014-04)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Ikeda, M., Tsay, J., and Weidman, L. (2012). “Exploratory Analysis of the Differences in American Community Survey Respondent Characteristics Between the Mandatory and Voluntary Response Methods,” *Research Report Series (Statistics #2012-01)*, Center for Statistical Research & Methodology, US Census Bureau, Wash. DC.
- Joyce, P., Malec, D., Little, R., Gilary, A., Navarro, A., and Asiala, M. (2014). “Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations,” *Journal of American Statistical Association*, 109 (505), 36-47.
- Klein, M. and Wright, T. (2011). “Ranking Procedures for Several Normal Populations: An Empirical Investigation,” *International Journal of Statistical Sciences, Volume 11 (P.C. Mahalanobis Memorial Special Issue)*, 37-58.
- Lu, B. and Ashmead, R. (In Press). “Propensity Score Matching Analysis for Causal Effect with MNAR Covariates,” *Statistica Sinica*.
- Mulry, M. H. (2014). “Measuring Undercounts in Hard-to-Survey Groups,” in R. Tourangeau, N. Bates, B. Edwards, T. Johnson, and K. Wolter (Eds.), *Hard-to-Survey Populations*. Cambridge University Press, Cambridge, England.
- Mulry, M., Kaputa, S. and Thompson, K. (In Press), “Initial M-estimation Parameter Settings for Detection and Treatment of Influential Values,” *Journal of Official Statistics*.
- Mary H. Mulry and Keller, A. (2017). “Comparison of 2010 Census Nonresponse Followup Proxy Responses with Administrative Records Using Census Coverage Measurement Results.” *Journal of Official Statistics*. 33(2). 455–475. DOI: <https://doi.org/10.1515/jos-2017-0022>
- Mary H. Mulry, Nichols, E. M., and Hunter Childs, J. (2017). “Using administrative records data at the U.S. Census Bureau: Lessons learned from two research projects evaluating survey data.” In Biemer, P.P, Eckman, S., Edwards, B., Lyberg, L., Tucker, C., de Leeuw, E., Kreuter, F., and West, B.T. *Total Survey Error in Practice*. Wiley. New York. 467-473.
- Mulry, M. H., Nichols, E. M., and Childs, J. Hunter (2016). “A Case Study of Error in Survey Reports of Move Month Using the U.S. Postal Service Change of Address Records,” *Survey Methods: Insights from the Field*. Retrieved from <http://surveyinsights.org/?p=7794>
- Mulry, M. H., Oliver, B. E., and Kaputa, S. J. (2014) “Detecting and Treating Verified Influential Values in a Monthly Retail Trade Survey.” *Journal of Official Statistics*, 30(4), 1–28.
- Mulry, Mary H., Oliver, B., Kaputa, S., and Thompson, K. J. (2016). “Cautionary Note on Clark Winsorization.” *Survey Methodology* 42 (2), 297-305. <http://www.statcan.gc.ca/pub/12-001-x/2016002/article/14676-eng.pdf>
- Nagaraja, C. and McElroy, T. (2015). “On the Interpretation of Multi-Year Estimates of the American Community Survey as

- Period Estimates.” Published online, *Journal of the International Association of Official Statistics*.
- Shao, J., Slud, E., Cheng, Y., Wang, S. and Hogue, C. (2014). “Theoretical and Empirical Properties of Model Assisted Decision-Based Regression Estimators,” *Survey Methodology* 40(1), 81-104.
- Slud, Eric. (2015). “Impact of Mode-based Imputation on ACS Estimates,” *American Community Survey Research and Evaluation Memorandum, #ACS-RER-O7*.
- Slud, E. and Ashmead, R., (2017), “Hybrid BRR and Parametric-Bootstrap Variance Estimates for Small Domains in Large Surveys”, *Proceedings of Survey Research Methods Section*, American Statistical Association, Alexandria, VA.
- Slud, E., Grievess, C. and Rottach, R. (2013). “Single Stage Generalized Raking Weight Adjustment in the Current Population Survey,” *Proceedings of Survey Research Methods Section*, American Statistical Association, Alexandria, VA.
- Slud, E. and Thibaudeau, Y. (2010). “Simultaneous Calibration and Nonresponse Adjustment,” *Research Report Series (Statistics#2010-03)*, Statistical Research Division, U.S. Census Bureau, Washington, DC.
- Thibaudeau, Y., Slud, E. and Gottschalck, A. (2017), “Modeling Log-linear Conditional Probabilities for Estimation in Surveys”, *Annals of Applied Statistics*, 11 (2), 680-697.
- Wieczorek, J. (2017). “Ranking Project: The Ranking Project: Visualizations for Comparing Populations,” R package version 0.1.1. URL: <https://cran.r-project.org/package=RankingProject>.
- Wright, T. (2012). “The Equivalence of Neyman Optimum Allocation for Sampling and Equal Proportions for Apportioning the U.S. House of Representatives,” *The American Statistician*, 66 (4), 217-224.
- Wright, T. (2013). “A Visual Proof, a Test, and an Extension of a Simple Tool for Comparing Competing Estimates,” *Research Report Series (Statistics #2013-05)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC.
- Wright, T., Klein, M., and Wieczorek, J. (2013). “An Overview of Some Concepts for Potential Use in Ranking Populations Based on Sample Survey Data,” *2013 Proceedings of the World Congress of Statistics (Hong Kong)*, International Statistical Institute.
- Wright, T. (2014). “A Simple Method of Exact Optimal Sample Allocation under Stratification with Any Mixed Constraint Patterns,” *Research Report Series (Statistics #2014-07)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC.
- Wright, T. (2014). “Lagrange’s Identity and Congressional Apportionment,” *The American Mathematical Monthly*, 121, 523-528.
- Wright, T. (2016). “Two Optimal Exact Sample Allocation Algorithms: Sampling Variance Decomposition Is Key,” *Research Report Series (Statistics #2016-03)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC.
- Wright, T. (2017). “Exact Optimal Sample Allocation: More Efficient Than Neyman,” *Statistics and Probability Letters*, 129, 50-57.
- Wright, T., Klein, M., and Wieczorek, J. (In Press). “A Primer on Visualizations for Comparing Populations, Including the Issue of Overlapping Confidence Intervals,” *The American Statistician*.

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

Funding Sources for FY 2018: 0331 – Working Capital Fund / 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.
- Diagnostics of seasonality must address differing sampling frequencies (monthly versus quarterly) and multiple forms of seasonality (cycles of annual versus weekly period), and must distinguish between raw and seasonally adjusted data.
- Multivariate modeling can not only provide increased precision of seasonal adjustments, but can also assist with series that have a low signal content. Moreover, aggregation constraints arising from accounting rules can be more readily enforced. This motivates the need to develop a viable multivariate seasonal adjustment methodology that can handle modeling, fitting, and seasonal adjustment of a large number of series.
- Time series data are being measured at higher sampling rates or over geographical regions, requiring new seasonal adjustment methods for high frequency/space-time data.

Potential Applications

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

Accomplishments (October 2016 – September 2017):

- Developed new estimation methods for vector time series models, allowing for parameter constraints and enforcement of stability.
- Developed new algorithms for forecasting and signal extraction of multivariate time series.
- Developed new tests of co-integration for multivariate time series.
- Investigated the causation of residual seasonality in indirect seasonal adjustments of time series, and developed a methodology for reconciliation of direct adjustments.
- Studied the modeling of seasonal time series measured at higher sampling frequencies (e.g., weekly and daily), and the implications for understanding moving holidays and trading day effects.
- Completed metrics for measuring and comparing the entropy of time series model residuals.
- Developed new models and filtering methods for extracting business cycles, trends, and seasonality from economic time series.

Short-Term Activities (FY 2018):

- Study and investigate diagnostics for seasonality, including the development of new diagnostics.
- Continue the study of high-frequency time series, including the facets of modeling, fitting, computation, separation of low-frequency signals, identification of holiday effects, attenuating of extremes, and applications to change of support problems.
- Finish studying the impact of weather on seasonal patterns and regARIMA models.
- Finish a comparison study of X-11 and SEATS seasonal adjustment, with respect to the mean square error of simulated seasonal adjustments.
- Generate an R package for *sigex*, and continue the dissemination of *x13story*.

Longer-Term Activities (beyond FY 2018):

- Further develop methods for constrained signal extraction, appropriate for multivariate data subject to accounting relations.
- Continue investigation of Seasonal Vector Form, allowing for more exotic seasonal models, and develop the corresponding seasonal adjustment methods.

- Expand research on multivariate seasonal adjustment in order to address the facets of co-integration, batch identification, modeling, estimation, and algorithms.
- Continue the development of X-13ARIMA-SEATS in order to improve the speed and stability of likelihood optimization and re-engineer the software around a more modern computing paradigm.
- 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.
- Investigate the properties and applications of both integer time series and network time series models.
- Develop and disseminate software to implement state space models, with the intention of incorporating sampling error and stochastic trading day into a seasonal adjustment framework.
- Develop estimators for the duration of a moving holiday effect.
- Continue investigation of cycles, band-pass filters, and signal extraction machinery for a broad array of signals.

Selected Publications:

- Alexandrov, T., Bianconcini, S., Dagum, E., Maass, P., and McElroy, T. (2012). "The Review of Some Modern Approaches to the Problem of Trend Extraction," *Econometric Reviews* 31, 593-624.
- Bell, W., Holan, S., and McElroy, T. (2012). *Economic Time Series: Modeling and Seasonality*. New York: Chapman Hall.
- Blakely, C. (2012). "Extracting Intrinsic Modes in Stationary and Nonstationary Time Series Using Reproducing Kernels and Quadratic Programming," *International Journal of Computational Methods*, Vol. 8, No. 3.
- Blakely, C. and McElroy, T. (2016). "Signal Extraction Goodness-of-fit Diagnostic Tests Under Model Parameter Uncertainty: Formulations and Empirical Evaluation," Published online, *Econometric Reviews*, 1-16.
- Findley, D. F. (2013). "Model-Based Seasonal Adjustment Made Concrete with the First Order Seasonal Autoregressive Model," Center for Statistical Research & Methodology, *Research Report Series (Statistics #2013-04)*. U.S. Census Bureau, Washington, DC.
- Findley, D. F., Monsell, B. C., and Hou, C.-T. (2012). "Stock Series Holiday Regressors Generated from Flow Series Holiday Regressors," *Taiwan Economic Forecast and Policy*.
- Holan, S. and McElroy, T. (2012). "On the Seasonal Adjustment of Long Memory Time Series," in *Economic Time Series: Modeling and Seasonality*. Chapman-Hall.
- Jach, A., McElroy, T., and Politis, D. (2012). "Subsampling Inference for the Mean of Heavy-tailed Long Memory Time Series," *Journal of Time Series Analysis*, 33, 96-111.
- Janicki, R. and McElroy, T. (2016). "Hermite Expansion and Estimation of Monotonic Transformations of Gaussian Data," *Journal of Nonparametric Statistics*, 28(1), 207-234.
- Lund, R., Holan, S., and Livsey, J. (2015). "Long Memory Discrete-Valued Time Series." Forthcoming, *Handbook of Discrete-Valued Time Series*. Eds R. Davis, S. Holan, R. Lund, N. Ravishanker. CRC Press.
- Lund, R. and Livsey, J. (2015). "Renewal Based Count Time Series." Forthcoming, *Handbook of Discrete-Valued Time Series*. Eds R. Davis, S. Holan, R. Lund, N. Ravishanker. CRC Press.
- McElroy, T. (2016). "Multivariate Seasonal Adjustment, Economic Identities, and Seasonal Taxonomy," Published online, *Journal of Business and Economics Statistics*.
- McElroy, T. (2016). "On the Measurement and Treatment of Extremes in Time Series," *Extremes*, 19(3), 467-490.
- McElroy, T. (2015). "When are Direct Multi-Step and Iterative Forecasts Identical?" *Journal of Forecasting*, 34, 315--336.
- McElroy, T. (2013). "Forecasting CARIMA Processes with Applications to Signal Extraction," *Annals of the Institute of Statistical Mathematics*, 65, 439-456.
- McElroy, T. (2012). "The Perils of Inferring Serial Dependence from Sample Autocorrelation of Moving Average Series," *Statistics and Probability Letters*, 82, 1632-1636.
- McElroy, T. (2012). "An Alternative Model-based Seasonal Adjustment that Reduces Over-Adjustment," *Taiwan Economic Forecast and Policy* 43, 33-70.
- McElroy, T. and Findley, D. (2015). "Fitting Constrained Vector Autoregression Models," in *Empirical Economic and Financial Research*.
- McElroy, T. and Holan, S. (2014) "Asymptotic Theory of Cepstral Random Fields," *Annals of Statistics*, 42, 64-86.
- McElroy, T. and Holan, S. (2012). "A Conversation with David Findley," *Statistical Science*, 27, 594-606.
- McElroy, T. and Holan, S. (2012). "On the Computation of Autocovariances for Generalized Geganbauer Processes," *Statistica Sinica* 22, 1661-1687.
- McElroy, T. and Holan, S. (2012). "The Error in Business Cycle Estimates Obtained from Seasonally Adjusted Data," in *Economic Time Series: Modeling and Seasonality*. Chapman-Hall.
- McElroy, T. and Jach, A. (2012). "Subsampling inference for the autocovariances of heavy-tailed long-memory time series," *Journal of Time Series Analysis*, 33, 935-953.
- McElroy, T. and Jach, A. (2012). "Tail Index Estimation in the Presence of Long Memory Dynamics," *Computational Statistics and Data Analysis*, 56, 266-282.

- McElroy, T. and Maravall, A. (2014). "Optimal Signal Extraction with Correlated Components," *Journal of Time Series Econometrics*, 6, 237--273.
- McElroy, T. and McCracken, M. (2015). "Multi-Step Ahead Forecasting of Vector Time Series." Published online, *Econometrics Reviews*.
- McElroy, T. and Monsell, B. (2014). "The Multiple Testing Problem for Box-Pierce Statistics," *Electronic Journal of Statistics*, 8, 497-522.
- McElroy, T. and Monsell, B. (2015). "Model Estimation, Prediction, and Signal Extraction for Nonstationary Stock and Flow Time Series Observed at Mixed Frequencies." *Journal of the American Statistical Association (Theory and Methods)*, 110, 1284-1303.
- McElroy, T. and Nagaraja, C. (2016). "Tail Index Estimation with a Fixed Tuning Parameter Fraction," *Journal of Statistical Planning and Inference*, 170, 27-45.
- McElroy, T. and Pang, O. (2015). "The Algebraic Structure of Transformed Time Series," in *Empirical Economic and Financial Research*.
- McElroy, T. and Politis, D. (2014). "Spectral Density and Spectral Distribution Inference for Long Memory Time Series via Fixed-b Asymptotics," *Journal of Econometrics*, 182, 211-225.
- McElroy, T. and Politis, D. (2013). "Distribution Theory for the Studentized Mean for Long, Short, and Negative Memory Time Series," *Journal of Econometrics*.
- McElroy, T. and Politis, D. (2012). "Fixed-b Asymptotics for the Studentized Mean for Long and Negative Memory Time Series," *Econometric Theory*, 28, 471-481.
- McElroy, T. and Trimbur, T. (2015). "Signal Extraction for Nonstationary Multivariate Time Series with Illustrations for Trend Inflation." *Journal of Time Series Analysis* 36, 209--227. Also in "Finance and Economics Discussion Series," Federal Reserve Board. 2012-45. <http://www.federalreserve.gov/pubs/feds/2012/201245/201245abs.html>
- McElroy, T. and Wildi, M. (2013). "Multi-Step Ahead Estimation of Time Series Models," *International Journal of Forecasting* 29, 378-394.
- Monsell, B. C. (2014) "The Effect of Forecasting on X-11 Adjustment Filters," *2014 Proceedings American Statistical Association* [CD-ROM]: Alexandria, VA.
- Monsell, B. C. and Blakely, C. (2013). "X-13ARIMA-SEATS and iMetrica," *2013 Proceedings of the World Congress of Statistics (Hong Kong)*, International Statistical Institute.
- Quenneville, B. and Findley, D. F. (2012). "The Timing and Magnitude Relationships Between Month-to-Month Changes and Year-to-Year Changes That Make Comparing Them Difficult," *Taiwan Economic Forecast and Policy*, 43, 119-138.
- Roy, A., McElroy, T., and Linton, P. (2014). "Estimation of Causal Invertible VARMA Models," *Cornell University Library*, <http://arxiv.org/pdf/1406.4584.pdf>.
- Trimbur, T. and McElroy, T. (2016). "Signal Extraction for Nonstationary Time Series With Diverse Sampling Rules," Published online, *Journal of Time Series Econometrics*.
- Wildi, M. and McElroy, T. (2016). "Optimal Real-Time Filters for Linear Prediction Problems," *Journal of Time Series Econometrics*, 8(2), 155-192.

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

Funding Sources for FY 2018: 0331 – Working Capital Fund / 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. In particular, linear mixed effects models are ubiquitous at the Census Bureau through applications of small area estimation. Models can also identify errors in data, e.g. by computing valid tolerance bounds and flagging data outside the bounds for further review.

Research Problems:

- 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.
- 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.
- Develop predictors of random effects in mixed models based on fiducial quantities that can be used for point estimation. Fiducial inference has seen a revival in recent years as an alternative to maximum likelihood, restricted maximum likelihood, and Bayesian methods. Point estimation with fiducial quantities has not yet been addressed; an advantage of this approach is that explicit point estimators of unknown parameters would not be required.
- Construct rectangular nonparametric tolerance regions for multivariate data. Tolerance regions for multivariate data are usually elliptical in shape, but such regions cannot provide information on individual components of the measurement vector. However, such information can be obtained through rectangular tolerance regions.
- Investigate statistical methods for remote sensing data, such as multispectral and LIDAR images, and potential applications at the Census Bureau.

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. Embedded experiments can be used to evaluate the effectiveness of alternative contact strategies.
- Expanding the collection of experimental design procedures currently utilized with the American Community Survey.
- Fiducial predictors of random effects can be applied to mixed effects models such as those used in small area estimation.
- Rectangular tolerance regions can be applied to multivariate economic data and aid in the editing process by identifying observations that are outlying in one or more attributes and which subsequently should undergo further review.

Accomplishments (October 2016 – September 2017):

- Released an updated COMPoisonReg package on CRAN. This version supports both zero-inflated and standard COM-Poisson regression modeling.
- Developed a prototype R package for spatio-temporal change of support modeling. Public American Community Survey (ACS) data was used to demonstrate that estimates on non-standard geographies and lookback periods, i.e. not released by the Census Bureau, could be produced by data users.
- A report on the design and analysis of embedded experiments is under preparation. The report summarizes the design and analysis of such experiments under various sampling and experimental designs. Variance estimation is discussed for

comparing different treatments, and Census Bureau embedded experiments examples are discussed from the ACS and CPS.

- The theoretical development is complete for deriving multivariate rectangular tolerance regions. Simulations are now being carried out.

Short-Term Activities (FY 2018):

- Develop flexible longitudinal Conway-Maxwell-Poisson model that allows for more general distributional forms to account for random effects.
- Complete R package and companion paper for spatio-temporal change of support modeling.
- Complete the report on the design and analysis of embedded experiments.
- Investigate modeling methodologies to evaluate the effectiveness of in-office address canvassing versus in-field canvassing.

Longer-Term Activities (beyond FY 2018):

- Develop generalized/flexible spatial and time series models motivated by the Conway-Maxwell-Poisson distribution.
- Significant progress has been made recently on randomization-based causal inference for complex experiments; Ding (*Statistical Science*, 2017), Dasgupta, Pillai and Rubin (*Journal of the Royal Statistical Society, Series B*, 2015), Ding and Dasgupta (*Journal of the American Statistical Association*, 2016) and Mukerjee, Dasgupta and Rubin (*Journal of the American Statistical Association*, 2017). It is proposed to adopt these methodologies for analyzing complex embedded experiments, by taking into account the features of embedded experiments (for example, random interviewer effects and different sampling designs).

Selected Publications:

- Gamage, G., Mathew, T. and Weerahandi, S. (2013). "Generalized Prediction Intervals for BLUPs in Mixed Models," *Journal of Multivariate Analysis*, 120, 226-233.
- Heim, K. and Raim, A.M. (2016). Predicting coverage error on the Master Address File using spatial modeling methods at the block level. In *JSM Proceedings, Survey Research Methods Section*. Alexandria, VA: American Statistical Association.
- Klein, M., Mathew, T. and Sinha, B. K. (2014). "Likelihood Based Inference Under Noise Multiplication," *Thailand Statistician*, 12(1), pp.1-23. URL: <http://www.tci-thaijo.org/index.php/thaistat/article/view/34199/28686>
- Mathew, T. and Young, D. S. (2013). "Fiducial-Based Tolerance Intervals for Some Discrete Distributions," *Computational Statistics and Data Analysis*, 61, 38-49.
- Mathew, T., Menon, S., Perevozskaya, I. and Weerahandi, S. (2016). "Improved Prediction Intervals in Heteroscedastic Mixed-Effects Models," *Statistics & Probability Letters*, 114, 48-53.
- Morris, D.S., Sellers, K.F., and Menger, A. (2017) Fitting a Flexible Model for Longitudinal Count Data Using the NLMIXED Procedure, SAS Global Forum Proceedings Paper 202-2017, SAS Institute: Cary, NC.
- Raim, A.M. and Gargano, M.N. (2015). "Selection of predictors to model coverage errors in the Master Address File," *Research Report Series: Statistics #2015-04*, Center for Statistical Research and Methodology, U.S. Census Bureau.
- Raim, A.M. (2016). Informing maintenance to the U.S. Census Bureau's Master Address File with statistical decision theory. In *JSM Proceedings*, Government Statistics Section. Alexandria, VA: American Statistical Association.
- Sellers, K., Lotze, T., and Raim, A. (2017) COMPoisonReg: Conway-Maxwell-Poisson Regression, version 0.4.0, 0.4.1, <https://cran.r-project.org/web/packages/COMPoisonReg/index.html>
- Sellers, K.F., and Morris, D. (In Press). "Under-dispersion Models: Models That Are 'Under The Radar'", *Communications in Statistics – Theory and Methods*.
- Sellers, K.F., Morris, D.S., and Balakrishnan, N. (2016). "Bivariate Conway-Maswell-Poisson Distribution: Formulation, Properties, and Inference," *Journal of Multivariate Analysis*, 150:152-168.
- Sellers K.F., Morris D.S., Shmueli, G., and Zhu, L. (2017). "Reply: Models for Count Data (a response to a letter to the editor)," *The American Statistician*.
- Sellers, K.F. and Raim, A.M. (2016). "A flexible zero-inflated model to address data dispersion". *Computational Statistics and Data Analysis*, 99: 68-80.
- Sellers, K., Morris, D., Balakrishnan, N., Davenport, D. (2017) multicomp: Flexible Modeling of Multivariate Count Data via the multivariate Conway-Maxwell-Poisson distribution, <https://cran.r-project.org/web/packages/multicomp/index.html>
- Young, D.S. (2013). "Regression Tolerance Intervals," *Communications in Statistics – Simulation and Computation*, 42(9), 2040-2055.
- Young, D.S. (2014). "A procedure for approximate negative binomial tolerance intervals", *Journal of Statistical Computation and Simulation*, 84(2), pp.438-450. URL: <http://dx.doi.org/10.1080/00949655.2012.715649>
- Young, D. and Mathew, T. (2015). "Ratio Edits Based on Statistical Tolerance Intervals." *Journal of Official Statistics* 31, 77-100.

Young, D.S., Raim, A.M., and Johnson, N.R. (2017). "Zero-inflated modelling for characterizing coverage errors of extracts from the US Census Bureau's Master Address File". *Journal of the Royal Statistical Society: Series A*. 180(1):73-97.

Zhu, L., Sellers, K.F., Morris, D.S., and Shmueli, G. (2017) Bridging the Gap: A Generalized Stochastic Process For Count Data, *The American Statistician*, 71 (1): 71-80.

Zhu, L., Sellers, K., Morris, D., Shmueli, G.,and Davenport, D. (2017) cmpprocess: Flexible Modeling of Count Processes, <https://cran.r-project.org/web/packages/cmpprocess/index.html>

Contact: Andrew Raim, Thomas Mathew, Kimberly Sellers, Dan Weinberg, Robert Ashmead, Scott Holan (R&M)

Funding Sources for FY 2018: 0331 – Working Capital Fund / 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 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.
- 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 infusion and synthetic data 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 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 2016 – September 2017):

- Developed new methodology that uses the principle of sufficiency to create synthetic data whose distribution is identical to the distribution of the original data under the normal linear regression model.
- Further developed and refined several data visualization methods for comparing populations and determining if there is a statistically significant difference between pairs of population parameters; applied the methodology to American Community Survey data.
- Developed finite sample methodology for drawing inference based on multiply imputed synthetic data under the multiple linear regression model.
- Evaluated bootstrap confidence intervals for unknown population ranks using simulation and proposed new uncertainty measures for estimated ranks using bootstrap.
- Applied small area estimation methodology to compute state and county level estimates based on the Tobacco Use Supplement to the Current Population Survey.
- Developed an interactive application using R Shiny to visualize high dimensional synthetic data and associated metrics.

- Further developed methodology for modeling response propensity using data from the National Crime Victimization Survey Field Representatives.
- Refined, expanded, and further developed a realistic artificial population that can now be used to simulate Monthly Wholesale Trade Survey data for a period representative of over four years.

Short-Term Activities (FY 2018):

- Continue developing finite sample methodology for drawing inference based on multiply imputed synthetic data and extend to multivariate models.
- Evaluate properties of bootstrap-based uncertainty measures for unknown population ranks.
- Evaluate properties of synthetic data when data generating, imputation, and analysis models differ under multivariate models.
- Use the constructed artificial population to implement simulation studies to evaluate properties of model-based estimation procedures for the Monthly Wholesale Trade Survey and other similar surveys.
- Develop and refine visualizations for synthetic data in higher dimensions.
- Implement model selection and diagnostics for a small area model applied to the Tobacco Use Supplement of the Current Population Survey.
- Develop methodology for drawing inference based on singly imputed synthetic data.

Longer-Term Activities (beyond FY 2018):

- Develop methodology for analyzing singly and multiply imputed synthetic data under various realistic scenarios.
- Develop noise infusion methods for statistical disclosure control.
- Study ways of quantifying the privacy protection/data utility tradeoff in statistical disclosure control.
- Develop and study bootstrap methods for sample survey data.
- Create an environment for simulating complex aspects of economic/demographic surveys.
- Develop bootstrap and/or other methodology for quantifying uncertainty in statistical rankings, and refine visualizations.

Selected Publications:

- Moura, R., Klein, M., Coelho, C. and Sinha, B. (2017). "Inference for Multivariate Regression Model based on Synthetic Data Generated under Fixed-Posterior Predictive Sampling: Comparison with Plug-in Sampling," *REVSTAT – Statistical Journal*, 15(2): 155-186.
- Klein, M. and Datta, G. (2017). "Statistical Disclosure Control Via Sufficiency Under the Multiple Linear Regression Model," *Journal of Statistical Theory and Practice*.
- Klein, M., and Sinha, B. (2016). "Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models," *Journal of Privacy and Confidentiality*, 7: 43-98.
- Klein, M., and Sinha, B. (2015). "Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models," *Sankhya B: The Indian Journal of Statistics*, 77-B, 293-311.
- Klein, M., and Sinha, B. (2015). "Likelihood-Based Inference for Singly and Multiply Imputed Synthetic Data under a Normal Model," *Statistics and Probability Letters*, 105, 168-175.
- Klein, M., and Sinha, B. (2015). "Likelihood-Based Finite Sample Inference for Synthetic Data Based on Exponential Model," *Thailand Statistician: Journal of The Thai Statistical Association*, 13, 33-47.
- Wright, T., Klein, M., and Wieczorek, J. (2014). "Ranking Populations Based on Sample Survey Data," *Center for Statistical Research and Methodology, Research and Methodology Directorate Research Report Series (Statistics #2014-12)*. U.S. Census Bureau. Available online: <http://www.census.gov/srd/papers/pdf/trs2014-12.pdf>.
- Klein, M., Lineback, J.F., and Schafer, J. (2014). "Evaluating Imputation Techniques in the Monthly Wholesale Trade Survey," *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association.
- Klein, M., Mathew, T., and Sinha, B. (2014). "Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-normal Regression Samples," *Journal of Privacy and Confidentiality*, 6, 77-125.
- Klein, M., Mathew, T., and Sinha, B. (2014). "Likelihood Based Inference under Noise Multiplication," *Thailand Statistician: Journal of The Thai Statistical Association*, 12, 1-23.
- Wright, T., Klein, M., and Wieczorek, J. (2013). "An Overview of Some Concepts for Potential Use in Ranking Populations Based on Sample Survey Data," *The 59th International Statistical Institute World Statistics Congress*, Hong Kong, China.
- Klein, M. and Sinha, B. (2013). "Statistical Analysis of Noise Multiplied Data Using Multiple Imputation," *Journal of Official Statistics*, 29, 425-465.
- Klein, M. and Linton, P. (2013). "On a Comparison of Tests of Homogeneity of Binomial Proportions," *Journal of Statistical Theory and Applications*, 12, 208-224.

- Klein, M., Mathew, T., and Sinha, B. (2013). "A Comparison of Statistical Disclosure Control Methods: Multiple Imputation Versus Noise Multiplication." *Center for Statistical Research and Methodology, Research and Methodology Directorate Research Report Series (Statistics #2013-02)*. U.S. Census Bureau. Available online: <http://www.census.gov/srd/papers/pdf/rrs2013-02.pdf>.
- Shao, J., Klein, M., and Xu, J. (2012). "Imputation for Nonmonotone Nonresponse in the Survey of Industrial Research and Development," *Survey Methodology*, 38, 143-155.
- Klein, M. and Wright, T. (2011). "Ranking Procedures for Several Normal Populations: An Empirical Investigation," *International Journal of Statistical Sciences*, 11, 37-58.
- Klein, M. and Creecy, R. (2010). "Steps Toward Creating a Fully Synthetic Decennial Census Microdata File," *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association.

Contact: Martin Klein, Isaac Dompok, Brett Moran, Bimal Sinha

Funding Sources for FY 2018: 0331 – Working Capital Fund / General Research Project
Various Decennial, Demographic, and Economic Projects