STATISTICAL EXPERTISE & GENERAL RESEARCH TOPICS

CENTER FOR STATISTICAL RESEARCH & METHODOLOGY

Research & Methodology Directorate
U.S. Bureau of the Census
(FY 2021 – FY 2025)

TOTALLY OR PARTIALLY FUNDED BY

- **WORKING CAPITAL FUND / DRAL RESEARCH PROJECT**

SEPTEMBER 2020
(REVISED – January 2024)

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. With this update, we have modified the names for some of the areas of expertise to better reflect change and expertise/interest of current staff as well as changing Census Bureau needs. We have also added a new effort focusing on cross-cutting statistical general research priorities where we will form small study/working groups (not necessarily limited to folks in our center) where there seems to be overlap and a real need to look forward.

Expertise for Collaboration and Research[1] **Page**

[1] The 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 & Observational Data Modeling

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

Research Problems:

- Correct quantification of the reliability of estimates with imputed values, as variances can be substantially greater than that computed nominally. Methods for adjusting the variance to reflect the additional uncertainty created by the missing data.
- Simultaneous imputation of multiple survey variables to maintain joint properties, related to methods of evaluation of model-based imputation methods.
- Integrating editing and imputation of sample survey and census responses via Bayesian multiple imputation and synthetic data methods.
- Nonresponse adjustment and imputation using administrative records, based on propensity and/or multiple imputation models.
- Development of joint modeling and imputation of categorical variables using log-linear models for (sometimes sparse) contingency tables.
- Statistical modeling (e.g., latent class models) for combining sample survey, census and/or alternative source data.
- Statistical techniques (e.g., classification methods, multiple imputation models) for using alternative data sources to augment or replace actual data collection.

Current Subprojects:

- Data Editing and Imputation for Nonresponse (Thibaudeau, Morris, Shao)
- Imputation and Modeling using Observational/Alternative Data Sources (Morris, Thibaudeau)

Potential Applications:

- Modeling approaches for integrating Economic Census editing and imputation processing, and developing multiple synthetic industry-level Economic Census micro-data.
- Modeling approaches for using administrative records in lieu of or to supplement Decennial Census field visits due to imminent and future design decisions.
- Adapt survey questions in the American Community Survey based on modeling of administrative record quality.
- Produce multiply imputed, synthetic and/or composite estimates of more geographical granular and timely economic activity based on third party data.

Accomplishments (October 2018-September 2020):

- Researched, adapted, and implemented nonparametric Bayesian hierarchical models developed by Kim et al. (2017) for integrating Economic Census editing and imputation processing with developing multiple synthetic industry-level Economic Census micro-data that can be publicly shared in place of suppressed estimates.
- Collaborated in adapting an R package based on Kim et al. (2017) to be specifically tailored to edit and multiply impute Economic Census data, and documented specifications in a user's guide.
- Developed multiple synthetic generators to produce industry-level Economic Census micro-data.
- Collaborated to develop Bayesian multiple imputation models for using third party data to produce geographically granular and timely retail sales experimental estimates.
- Applied and completed evaluation of optimization methods for raking balance complexes in the Quarterly Financial Report (QFR) when items can take negative values.
- Showed how to use log-linear models coupled with complementary logistic regression to improve the efficiency (reducing the sampling error) of estimates of gross flows and month-to-month proportions classified by demographic variables. Illustrated methodology on labor force measurements and gross flows estimated from the Current Population Survey.

Short-Term Activities (FY 2021 - FY 2023):

- 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.
- Continue research, implementation, and resolution of editing and data issues when applying non-parametric Bayesian editing methods to edit and multiply impute Economic Census data.
- Continue research on integration of Bayesian editing and multiple imputation processing with disclosure avoidance and data synthesis processing.
- Extend the analysis and estimation of changes in the labor force status using log-linear models coupled with matching logistic regression methods to the Current Population Survey.
- Research novel categorical distributions for contingency table modeling and joint imputation of categorical variables particularly for clustered data.
- Continue research on accounting for observed zero cells in loglinear models for sparse contingency tables.

Longer-Term Activities (beyond FY 2023):

- Joint modeling of response propensity and administrative source accuracy.
- Research practical ways to apply decision theoretic concepts to the use of administrative records (versus personal contact or proxy response) in the Decennial Census.
- Further development of joint administrative record and imputation modeling based on latent class models.
- 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.
- Joint models of attrition (or response rate) and clustered categorical outcomes using shared random effects with innovative GLMM computational techniques.
- 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 in tandem with logistic regression.

Selected Publications:

Morris, D.S. and Raim, A.M. (2023). “Comparing Trial and Variable Association in Contingency Table Data Using Multinomial Models for Clustered Data.” *In Proceedings of the 37th International Workshop on Statistical Modelling*. Dortmund, Germany: Statistical Modelling Society, 536-542.

Kang, J., Morris, D.S., Joyce, P., and Dompreh, I. (In Press). “On Calibrated Inverse Probability Weighting and Generalized Boosting Propensity Score Models for Mean Estimation with Incomplete Survey Data,” *Wiley Interdisciplinary Reviews (WIREs) Computational Statistics*.

Morris, D.S. and Sellers, K.F. (2022). “A Flexible Mixed Model for Clustered Count Data,” *Stats: Special Issue on Statistics, Data Analytics, and Inferences for Discrete Data, 5(1):* 52–69. https://doi.org/10.3390/stats5010004.

Liu, B., Dompreh, I., and Hartman, A.M. (2021). “Small Area Estimation of Smoke-Free Workplace Policies and Home Rules in U.S. Counties,” *Journal of Nicotine and Tobacco Research*.

Dumbacher, B., Morris, D.S., and Hogue, C. (2019). “Using Electronic Transaction Data to Add Geographic Granularity to Official Estimates of Retail Sales,” *Journal of Big Data, 6(80)*.

Keller, A., Mule, V.T., Morris, D.S., and Konicki, S. (2018). "A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 Census," *Journal of Official Statistics, 34(3)*: 1-27.

Winkler, W. E. (2018). “Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Edit/Imputation,” *Research Report Series (Statistics #2018-05)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.

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.

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.

Thibaudeau, Y. and Morris, D.S. (2016). “Bayesian Decision Theory to Optimize the Use of Administrative Records in Census NRFU,” *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association.

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.*

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, D.C.

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 D.C.

Thibaudeau, Y. (2002). “Model Explicit Item Imputation for Demographic Categories,” *Survey Methodology, 28(2),* 135-143.

Contact: Darcy Morris, Joseph Kang, Shane Lubold, Isaac Dompreh, Yves Thibaudeau, Jun Shao, Eric Slud, Xiaoyun Lu

Funding Sources for FY 2021-2025: 0331 – Working Capital Fund / General Research Project
Various Decennial, Demographic, and Economic Projects

Record Linkage & Machine Learning

Motivation:

Record linkage continues to grow in importance as a fundamental activity in statistical agencies. The number of available administrative lists and commercial files has grown exponentially and present statistical agencies with opportunities to accumulate information through record-linkage to support the production of official statistics. In addition to cost, new obstacles to traditional data collection have emerged in the form of possibly recurrent pandemics. These circumstances further motivate the accumulation of information by linking public, private and administrative files. Thibaudeau (2020) describes the strides the Census Bureau, a pioneer in record linkage, has made over the years. While this is impressive, more is needed. With its own suite of in-house record-linkage software packages, such as the "SAS (PVS) Matcher," "BigMatch," "d-blink" and "MAMBA," and easy access to open-source packages, such as "fastLink" and "RecordLinkage in R," the Census Bureau now has access to a wide spectrum of methodologies and the potential to rapidly develop and integrate new ones. The Census Bureau must remain abreast of the ever improving state-of-the-art in record linkage and be prepared to champion its own methodologies as some of the best in the world. Our goal is to achieve the synergy of methods and software that will benefit most the Census Bureau and its mission. System portability is also an objective. The Census Bureau should have the freedom to upgrade its IT infrastructure knowing record-linkage applications will remain functional.

Research Problems:

One challenge is continuing to research and experiment with new methodologies on multiple software platforms while also moving toward integration. Description of such experiments are:

- Markov Chains Monte-Carlo (MCMC), like that powered by d-blink, give full probabilistic characterizations of the record-linkage process and are becoming indispensable for full comprehension of a record linkage process. At the same time MCMCs can be tweaked to deliver fast snapshots of the linked population. Research in that direction is crucial. Old-School programs like BigMatch have been greatly optimized for fast linking but lack in nuance. They need to be garnished by richer comparison schemes, such as dictionary-assisted fuzzy string comparisons.
- New data structure for record-linkage of multiple large lists need to be explored. d-blink is an example of a more efficient data structure: Node-connected structures minimize the number of comparisons, as opposed to a traditional all pairwise comparisons. Other structures are possible, such as cyclical linked lists (Thibaudeau 1992), and should be researched.
- As new techniques continue to be implemented and experimented on various existing software (R, Python, C) and hardware (Windows, OSX, IRE, CAES) platforms, the dominant paradigms are emerging and work toward integration and unification, while maintaining versatility, is moving in high gear.

Current Subprojects:

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

Potential Applications:

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

Accomplishments (October 2018-September 2020):

- Deployed the FEBRL (Peter Christen) Python file simulator on multiple platforms.
- Used FEBRL to simulate candidate files –up to 500k records each- for record-linkage with known "true-links." Used the simulated files to assess the rates of precision and recall of "BigMatch" and the "SAS (PVS) matcher."
- Presented poster at Data Linkage Day 2019 entitled: "False Duplicates in the Census: A Novel Approach to identifying False Matches from Record Linkage Software."
- Wrote EM algorithm for estimating the weights of the Fellegi-Sunter record-linkage model for use with "BigMatch" and the R "RecordLinkage" package.
- Installed BigMatch on multiple platforms IRE, Windows, MacOS (simulated data).

- Ran experiments to measure and compare the performance in speed and CPU cycles of a multi-core linkage software (R fastLink) and an optimized single-core record-linkage software (BigMatch) in a multi-core environment.
- Used BigMatch for multiple linkage projects, including the linkage of commercial files, in the construction of a master reference file at the person and housing unit levels for research and experimentation in preparation for Census 2030.

Short-Term Activities (FY 2021 - FY 2023):

- Provide advice to individuals who plan to update and maintain the programs for record linkage and related data preparation.
- Conduct research on record linkage error-rate estimation, particularly for unsupervised and semi-supervised situations.
- Evaluate "R" vs "Python" packages for record linkage focusing on fuzzy string comparison.
- Assess the possibility of using a surname and given-name reference directory for record-linkage in decennial-census production.
- Continue to research statistical and data-science methods for record linkage. Explore and compare in-house and "off-the-shelf" packages implementing these methods. Ascertain the competency of record-linkage methods at the Census Bureau.
- Extending record linkage outside the PIK universe.

Longer-Term Activities (beyond FY 2023):

- Construct census-based equivalence dictionaries of U.S. given names and surnames for cross-referencing and supervised learning in record-linkage.
- Integrate new methods in our in-house record-linkage engines. Consider the integration of off-the-shelf packages when advantageous.
- Evaluate and compare in-house and off-the-shelf data-science programs and packages (R and/or Python) to construct engines for massive numbers of record-linkage runs for Census 2030.
- Further develop Markov Chain Monte-Carlo applications embedding record-linkage methods in massive parallel processing. Develop methods for extracting record-linkage snapshots from MCMCs.

Selected Publications:

Wang, Z., Ben-David, E., and Slawski, M. (2023). "Regularization for Shuffled Data Problems via Exponential Family Priors on the Permutation Group." (*Proceedings of the 26th International Conference on Artificial Intelligence and Statistics), Proceedings of Machine Learning Research*, Volume 206, pgs 2939-2959. https://proceedings.mlr.press/v206/wang23a.

Steorts, R. (2023). "A Primer on the Data Cleaning Pipeline," *Journal of Survey Statistics and Methodology*, 11, 553-568.

Marchant, N.G., Rubinstein, B.I.P., and Steorts, R. (2023), "Bayesian Graphical Entity Resolution Using Exchangeable Random Partition Priors," *Journal of Survey Statistics and Methodology*, 11, 569-596.

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

Basak J., Soliman A., Deo N., Haase, K., Mathur, A., Park, K., Steorts, R., Weinberg, D., Sahni. S., and Sanguthevar R. (2023). "On Computing the Jaro Similarity Between Two Strings," *Proceedings of the 19th International Symposium on Bioinformatics Research and Applications*, Springer, 31-44.

Aleshin-Guendel, S. and Steorts, R. (In Press). "Monitoring Convergence Diagnostics for Entity Resolution," *Annual Review of Statistics and Its Applications*.

Betancourt, B., Zanella, G., and Steorts, R. (In Press). "Random Partition Models for Microclustering Tasks," *Journal of the American Statistical Association, Theory and Methods.*

Mosaferi, S., Ghosh, M., and Steorts, R. (In Press). "Measurement Error Models for Small Area Estimation," *Communications and Statistics: Simulation and Computation.*

Wang, Z., Ben-David, E., Diao, G., & Slawski, M. (In Press). "Estimation in Exponential Family Regression Based on Linked Data Contaminated by Mismatch Error," *Statistics and Its Interface*.

Wang, Z., Ben-David, E., Diao, G., & Slawski, M. (2022). "Regression with Linked Datasets Subject to Linkage Error," *Wiley Interdisciplinary Reviews: Computational Statistics, 14(4).*

Marchant, N., Kaplan, A., Rubenstein, B., Elzar, D., and Steorts, R. (2021). "d-blink: Distributed End-to-End Bayesian Entity Resolution," *Journal of Computational Graphics and Statistics*, *30(2),* 406-421.

Slawski, M., Diao, G., and Ben-David, E. (2021). "A Pseudo-Likelihood Approach to Linear Regression with Partially Shuffled Data," *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2020.1870482

Thibaudeau, Y., Slud, E., and Cheng, Y. (2021). "Small-Area Estimation of Cross-Classified Gross Flows Using Longitudinal Survey Data," *Advances in Longitudinal Survey Methodology*,469-489, Peter Lynn ed., Wiley.

Wang, Z., Ben-David, E., Diao, G., and Slawski, M. (2021). "Regression with Linked Datasets Subject to Linkage Error," Wiley Interdisciplinary Reviews: Computational Statistics, DOI: 10.1002/wics.1570

Thibaudeau, Y. (In progress). "New Record Linkage Solutions for Demographic Methods at the Census Bureau," Research Report Series (Statistics #2020-??), Center for Statistical Research & Methodology, U.S. Census Bureau, Washington, D.C.

Slawski, M. and Ben-David, E. (2019). "Linear Regression with Sparsely Permuted Data," *Electronic Journal of Statistics*, Vol 13, No. 1, 1-36.

Slud, E. and Thibaudeau, Y. (2019). "Multi-outcome Longitudinal Small Area Estimation – A Case Study," *Statistical Theory and Related Fields*, DOI: 10.1080/24754269.2019.1669360.

Steorts, R.J., Tancredi, A., and Liseo, B. (2018). "Generalized Bayesian Record Linkage and Regression with Exact Error Propagation" in Privacy in Statistical Databases (Lecture Notes in Computer Science 11126) (Eds.) Domingo-Ferrer, J. and Montes, F., Springer, 297-313.

Steorts, R.J. and Shrivastava, A. (2018). "Probabilistic Blocking with an Application to the Syrian Conflict," in Privacy in Statistical Databases (Lecture Notes in Computer Science 11126) (Eds.) Domingo-Ferrer, J. and Montes, F., Springer, 314-327.

Winkler, W. E. (2018). "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.

Thibaudeau, Y., Slud, E., and Gottshalck, A. (2017). "Log-Linear Conditional Probabilities for Estimation in Surveys," *Annals of Applied Statistics*, *11*, 680-697.

Czaja, W., Hafftka, A., Manning, B., and Weinberg, D. (2015). "Randomized Approximations of Operators and their Spectral Decomposition for Diffusion Based on Embeddings of Heterogeneous Data," 3rd International Workshop on Compressed Sensing Theory and Its Applications to Radar, Sonar and Remote Sensing (CoSeRa).

Winkler, W. E. (2015). "Probabilistic Linkage," in (H. Goldstein, K. Harron, C. Dibben, eds.) *Methodological Developments in Data Linkage*, J. Wiley: New York.

Weinberg, D. and Levy, D. (2014). "Modeling Selective Local Interactions with Memory: Motion on a 2D Lattice," *Physica D* 278-279, 13-30.

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. (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. (2011). "Machine Learning and Record Linkage" in *Proceedings of the 2011 International Statistical Institute*.

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. (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., 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.

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*.

Alvarez, M., Jonas, J., Winkler, W. E., and Wright, R. "Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project," *Electronic Voting Technology*.

Winkler, W. E. (2008). "Data Quality in Data Warehouses," in (J. Wang, Ed.) *Encyclopedia of Data Warehousing and Data Mining (2nd Edition)*.

Herzog, T. N., Scheuren, F., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*, New York, NY: Springer.

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, D.C.

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, D.C.

Winker, 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.

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, D.C.

Thibaudeau, Y. (2002). "Model Explicit Item Imputation for Demographic Categories," *Survey Methodology*, *28*, 135-143.

Thibaudeau, Y. (1993). "The Discrimination Power in Dependency Structure in Record Linkage," *Survey Methodology*, *19*, 31-38

Thibaudeau, Y. (1992). "Identifying Discriminatory Models in Record Linkage," Proceedings of the Section on Statistical

Computing, American Statistical Association, Alexandria, VA.

Winkler, W. and Thibaudeau, Y. (1991). “An Application of the Fellegi-Sunter Model of RecordLinkage to the 1990 Decennial Census,” *Research Report Series (Statistics) RR91/09*, Statistical Research Division, U.S. Census Bureau, Washington, D.C.

Contact: Yves Thibaudeau, Edward H. Porter, Emanuel Ben-David, Rebecca Steorts, Dan Weinberg

Funding Sources for FY 2021-2025: 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 of models that combine data across multiple surveys or combines survey and observational data (non-probability samples) to improve survey estimates.
- Development of model diagnostic and model comparison tools for small area models.
- Development of small area share models for subareas estimates (e.g., school districts or tracts).
- Development of a design-based simulation system which mimics the American Community Survey to use as a test-bed for area- and unit-level small area models, estimation (both model-based and design-based) methodology and estimation of uncertainty measures.
- Study of measurement error in small area estimation models.
- Development of temporal small area estimation techniques.
- Development of spatial small area estimation techniques.
- Development of more robust estimates of mean squared error of prediction by incorporating Bayesian and bootstrap methods.
- Development of unit-level model framework which appropriately takes into account the complex design of the survey.

Current Subprojects:

- Using ACS Estimates to Improve Estimates from Smaller Surveys via Bivariate Small Area Estimation Models (Franco, Bell/R&M)
- Bootstrap Mean Squared Error Estimation for Small Area Means under Non-normal Random Effects (Maples, Datta, Irimata, Slud)
- Developing correlated small area share models to create estimates of school district child poverty and population (Maples)
- Developing graphical methods to assess the assumption of constant parameter values across all domains (Maples, Dompreh)
- Developing Bayesian pseudolikelihood models for unit-level data obtained from a complex sample survey (Janicki)
- Assessment of mean squared errors of empirical best linear unbiased predictors for misspecified models (Datta, Slud)

Potential Applications:

- Borrowing strength from ACS estimates using bivariate modeling has many potential applications, including improving estimates from smaller surveys such as SIPP, NHIS, and CPS, and improving the ACS one year estimates themselves using the previous ACS 5-year estimates.
- Model diagnostic and comparison tools can be applied in any small area application, from SAIPE to SAHIE, to small area models applied to SIPP, AHS, etc.
- The design-based simulation framework for evaluating modes can be used for SAIPE, SAHIE, and other small area programs that use ACS data. The framework can also test the properties of design-based/assisted estimation procedures, such as improvements of sampling variance estimates, propensity score models etc.
- Temporal extensions of small area models will be potentially useful in the VRA Section 203B determinations, and can be applied to ACS data in general, as well as to other surveys that are repeated over time.
- The evaluation of measurement error will help determine if it is appropriate to use ACS-estimates as covariates in models for the Section 203B determinations, and at what level of aggregation.
- Small area share models may be a replacement to the current for the current school district estimates procedures for SAIPE.
- Spatial small area models can improve estimates and provide limited disclosure avoidance for some of the ACS special tabulations.

Accomplishments (October 2018-September 2020):

- Developed empirical and theoretical evidence that shows the strong potential of borrowing strength from ACS estimates to improve estimates from smaller U.S. sample surveys using simple bivariate small area estimation models, including applications to NHIS and SIPP, and an application that improves ACS one-year estimates using previous five-year estimates.
- Developed a small area share model to estimate the number of school aged children in poverty for school districts given the

official county level poverty estimates.

- Studied alternative models for SAIPE county estimates of school-aged children in poverty using a design-based simulation, and explored the impact of sampling variance estimation in model selection, exploring how design-based estimate and GVF-based estimates impact performance.
- Derived several different mean squared error estimators, both analytical and bootstrapped-based, which will be evaluated in a large simulation study.
- Studied the impact of differential privacy noise infusion on voting district plans and evaluated measures of variability.

Short-Term Activities (FY 2021 – FY 2023):

- Extend the Small Area Shares model to allow for dependence between sets of shares, e.g., allow the school district to county shares of school age children in poverty and not-in-poverty to have a dependence.
- Finish creating the Artificial Population which mimics the distribution of the U.S. population and implement an ACS-like survey design.
- Improve predictions in ACS special tabulations using a mixture of spatial models.
- Evaluate different mean squared error estimates under the Fay-Herriot model when the error distribution is not always correctly specified.
- Study the impact of measurement error in covariates in small area models for the Voting Rights Act Section 203 determinations.
- Explore times series extensions of the Multinomial Logit model and determine suitability for Voting Rights Act Section 203 determinations.
- Develop multivariate spatial models which use differentially private measurements and auxiliary survey data for the purpose of predicting the number of persons in counties and AIAN areas for detailed race groups.

Longer-Term Activities (beyond FY 2023):

- Develop graphical methods to test assumptions about constant model parameters across all areas.
- Develop models that jointly model survey-weighted proportions and effective sample sizes.
- Explore if a time series model can be applied to improve sampling variance estimates by borrowing strength from estimates from previous years.
- Evaluation of new models (county and school district) to update official SAIPE methodology.
- Deliver a set of 1000 independent survey samples from the Artificial Population with a design similar to the American Community Survey.

Selected Publications:

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

Datta, G.S., Lee, J., and Li, J. (In Press). "Pseudo-Bayesian Small Area Estimation," *Journal of Survey Statistics and Methodology*.

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

Parker, P.A., Janicki, R., and Holan, S. (In Press). "Bayesian Methods Applied to Small Area Estimation for Establishment Statistics," in Bavdaž, M., Bender, S., Jones, J., MacFeely, S., Sakshaug, J.W., Thompson, K.J., and van Delden, A. (Eds.), *Advances in Business Statistics, Methods and Data Collection*, Wiley.

Parker, P., Holan, S., and Janicki, R. (2022). "Computationally Efficient Bayesian Unit-level Models for Non-Gaussian Data Under Informative Sampling with Application to Estimation of Health Insurance Coverage," *The Annals of Applied Statistics, Vol 16, No. 2,* 887-904.

Ghosh, T., Ghosh, M., Maples, J., and Tang, X. (2022). "Multivariate Global-Local Priors for Small Area Estimation," *STATS, v5*, 673-688. https://www.mdpi.com/2571-905X/5/3/40/htm.

Janicki, R., Raim, A.M., Holan, S.H., and Maples, J. (2022). "Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models with Application to American Community Survey Special Tabulations," *The Annals of Applied Statistics, Volume 16, Issue 1*, 144-168.

Erciulescu, A., Franco, C., and Lahiri, P. (2021). "Use of Administrative Records in Small Area Estimation," in Chun, A. Y. and Larsen, M. (Eds.), *Administrative Records for Survey Methodology,* New York, NY: Wiley Publishers.

Liu, B., Dompreh, I., and Hartman, A.M. (2021). "Small Area Estimation of Smoke-Free Workplace Policies and Home Rules in U.S. Counties," *Journal of Nicotine and Tobacco Research*.

Parker, P. A., Holan, S. H., and Janicki, R. (2020). "Bayesian Unit-Level Modeling of Count Data under Informative Sampling Designs," Stat, 9.

Maples, J. (2019). "Small Area Estimates of the Child Population and Poverty in School Districts Using Dirichlet-Multinomial Models," *2019 Proceedings of the American Statistical Association,* Section on Survey Research Methods, American Statistical Association, Alexandria, VA, 3150-3152.

Bell, W. R., Chung, H. C., Datta, G. S., and Franco, C. (2019). "Measurement Error in Small Area Estimation: Functional vs.

Structural vs. Naïve Models," *Survey Methodology, 45,* 61-80.

Chakraborty, A., Datta, G.S., and Mandal, A. (2019). "Robust Hierarchical Bayes Small Area Estimation for Nested Error Regression Model," *International Statistical Review, 87, S1,* S158–S176, doi:10.1111/insr.12283.

Chung, H., Datta, G., and Maples, J. (2019). "Estimation of Median Incomes of the American States: Bayesian Estimation of Means of Subpopulations," *Opportunities and Challenges in Development*, Simanti Bandyopadhyay and Mousumi Datta (ed.), New York: Springer, 505-518.

Franco, C., Little, R. J. A., Louis, T. A., and Slud, E. V. (2019). "Comparative Study of Confidence Intervals for Proportions in Complex Surveys," *Journal of Survey Statistics and Methodology,* 7, 3, 334-364.

Datta, G.S., Rao, J.N.K., Torabi, M., and Liu, B. (2018). "Small Area Estimation with Multiple Covariates Measured with Errors: A Nested Error Linear Regression Approach of Combining Two Surveys," *Journal of Multivariate Analysis, 167,* 49-59.

Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2017). "Multivariate Fay-Herriot Bayesian Estimation of Small Area Means Under Functional Measurement Error," *Journal of the Royal Statistical Society--Series A, 180(4),* 1191-1209.

Janicki, R. and Vesper, A. (2017). "Benchmarking Techniques for Reconciling Small Area Models at Distinct Geographic Levels," *Statistical Methods Applications,* DOI: https://doi.org/10.1007/s10260-017-0379-x, *26*, 557-581.

Maples, J. (2017). "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, 180(4),* 1211-1227.

Chakraborty, A., Datta, G.S., and Mandal, A. (2016). "A Two-component Normal Mixture Alternative to the Fay-Herriot Model," *Joint issue of Statistics in Transition new series and Survey Methodology, Part II, 17*, 67-90.

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, D.C.

Datta, G.S. and Mandal, A. (2015). "Small Area Estimation with Uncertain Random Effects," *Journal of the American Statistical Association: Theory and Methods, 110,* 1735-1744.

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.

Bell, W.R., Datta, G.S., and Ghosh, M. (2013). "Benchmarking Small area Estimators," *Biometrika, 100,* 189-202, doi:10.1093/biomet/ass063.

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.

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.

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.

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, D.C.

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.

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, D.C.

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.

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.

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, D.C.

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.

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.

Contact: Jerry Maples, Ryan Janicki, Gauri Datta, Kyle Irimata, Bill Bell (ADRM), Eric Slud

Funding Sources for FY 2021-2025: 0331 – Working Capital Fund / General Research Project
Various Decennial, Demographic, and Economic Projects

Spatial Analysis & Modeling

Motivation:

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

Research Problems:

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

Current Subprojects:

- Spatio-temporal methods for simultaneous shrinkage of both means and variances for small area estimation. (Holan, Janicki, Parker)
- Developing Bayesian pseudolikelihood models for unit-level data obtained from a complex sample survey. (Holan, Janicki, Parker)
- Development of unit-level models with temporal dependence. (Holan, Janicki, Parker)
- Development of change-of-support methodology for inference on regions with no direct measurement, based on observations on a distinct geographic region or grid. (Holan, Janicki)
- Incorporation of spatially-correlated random effects in small area models. (Datta, Janicki, Maples)
- Development of model-based methods for improving survey estimates at low geographic levels, such as tract and block group. (Holan, Janicki, Parker)
- Accurate measurement of the uncertainty associated with predictions from highly-complex models. (Holan, Janicki, Parker)
- Integration of deep learning with spatial modeling. (Holan, Janicki, Parker)
- Obtaining consistency results when observations are irregularly spaced. (Lahiri)
- Generation of synthetic micro data from complex spatio-temporal models which preserves properties and dependencies found in the original data and can be published without disclosing confidential information. (Holan, Janicki, Parker)

Potential Applications:

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

Accomplishments (October 2018-September 2020):

- Developed a multivariate spatial mixture model for American Community Survey special tabulations which can be used to produce model-based predictions when the survey-specific sample size is insufficient, either due to privacy concerns or data quality concerns.
- Developed spatial models for differentially private measurements of decennial census counts and ratios for improving precision and aggregating to marginal table cells.

- Developed a spatial change-of-support model for predicting counts in regions where no direct response variable is available.

Short-Term Activities (FY 2021 – FY 2023):

- Produce model-based estimates of 2010 decennial census counts using spatial models fit to differentially private measurements for nine target table shells.
- Exploration of novel uses of auxiliary data and data integration for im- proved prediction and development of new data products.
- Research the extent to which utilization of spatial information and multivariate dependencies can reduce the impact of the effect of differential privacy on the precision of data products.
- Development of software for efficiently fitting a variety of spatial, spatio-temporal, longitudinal, mixture, and other hierarchical Bayesian models.
- Investigate new and efficient computational methods for fitting high-dimensional models.

Longer-Term Activities (beyond FY 2023):

- Development of model-based methods for inference on very small domains, such as block groups, when the data are very sparse and are of sufficient quality for publication.
- Development of efficient methods for producing special tabulations which of survey data and which meet the U. S. Census Bureau's data quality standards.
- Development of methodology for producing estimates at non-standard geographies such as American Indian and Alaska Native areas and school districts
- Methodology for producing synthetic microdata which can be made publicly available for data users.

Selected Publications:

Parker, P., Holan, S.H., and Janicki, R. (2023). "Comparison of Unit Level Small Area Estimation Modeling Approaches for Survey Data Under Informative Sampling," *Journal of Survey Statistics and Methodology*, Vol 11, No. 4, 858-872.

Parker, P., Holan, S.H., and Janicki, R. (2023). "A Comprehensive Overview of Unit Level Modeling of Survey Data for Small Area Estimation Under Informative Sampling," *Journal of Survey Statistics and Methodology*, Vol 11, No. 4, 829-857.

Parker, P., Holan, S.H., and Janicki, R. (In Press). "Conjugate Modeling Approaches for Small Area Estimation with Heteroscedastic Structure," *Journal of Survey Statistics and Methodology*.

Janicki, R., Holan, S. H., Irimata, K. M., Livsey, J., and Raim, A. (In Press). "Spatial Change of Support Models for Differentially Private Decennial Census Counts of Persons by Detailed Race and Ethnicity," *Journal of Statistical Theory and Practice*.

Parker, P., Holan, S. H., and Janicki, R. (2022). "Computationally Efficient Bayesian Unit-Level Models for Multivariate Non-Gaussian Data Under Informative Sampling." *Annals of Applied Statistics*, 16, 887 – 904.

Janicki, R., Raim, A., Holan, S. H., and Maples, J. (2022). "Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models with Application to American Community Survey Special Tabulations." *Annals of Applied Statistics*, 16, 144 – 168.

Parker, P., Holan, S. H., and Janicki, R. (2020). "Conjugate Bayesian unit-level modeling of count data under informative sampling designs." *Stat*, 9, e267.

Contact: Ryan Janicki, Soumen Lahiri, Paul Parker, Scott Holan (ADRM), Serge Aleshin-Guendel

Funding Sources for FY 2021-2025: 0331 – Working Capital Fund / General Research Project
Various Decennial, Demographic, and Economic Projects

Sampling Estimation & Survey Inference

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 inclusion in observational or administrative lists be modeled jointly with indicator and mode of survey response, so that traditional survey methods can be extended to merged survey and non-survey data?
- Can non-traditional design methods such as adaptive sampling be used to improve estimation for rare characteristics and populations?
- How can time series and spatial methods be used to improve ACS estimates or explain patterns in the data?
- Can generalized weighting methods be formulated and solved as optimization problems to avoid the ambiguities resulting from multiple weighting step and to explicitly allow inexact calibration?
- What models can aid in assessing the combined effect of all the sources of sampling and nonsampling error, including frame coverage errors and measurement errors, on sample survey estimates?
- What experiments and analyses can inform the development of outreach methods to enhance census response?
- Can unduplication and matching errors be accounted for in modeling frame coverage in censuses and sample surveys?
- How can small-area or other model-based methods be used to improve interval estimates in sample surveys, to design survey collection methods with lowered costs, or to improve Census Bureau imputation methods?
- Can classical methods in nonparametrics (e.g., using ranks) improve estimates from sample surveys?
- How can we measure and present uncertainty in rankings of units based on sample survey estimates?
- Can Big Data improve results from censuses and sample surveys?
- How to develop and use bootstrap methods for expressing uncertainty in estimates from probability sampling?

Current Subprojects:

- Optimization-based (single-stage) approaches to Weight-adjustment for Probability and Nonprobability Samples (Slud, Morris)
- The Ranking Project: Methodology Development and Evaluation (Wright,Klein/FDA, Wieczorek/Colby College, Yau)
- Optimal Sample Allocation and Apportionment (Wright)
- Optimal stratification in economic surveys, using multiple measures of size and multiple survey outcomes (Slud, Joyce)
- Machine Learning projects related to non-response segmentation Mindsets for decennial outreach (Mulry, Morris, Scheid/DSSD), or to Frames (Weinberg, Slud)
- Methods of estimating variances for survey estimates combining model- and design-based estimates, and simulation studies of bias when the design-based methods include Replication Methods in domains with small sample-size (Slud, Trudell)
- Analyses supporting improvement of household rosters for census nonresponders that are projected to be occupied and to have high quality administrative records. (Mulry).

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, 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 2018-September 2020):

- Contributed to team development of methods for producing differentially private decennial census tabulations conforming to legally mandated error-free disclosure of block-level population totals under *Public Law 94* as well as to *Title 13* requirements for nondisclosure of individual-level data.
- Developed novel optimization-based weighting adjustment methods based on partially missing data, along with diagnostics based on cross-classified post-stratification variables.
- Demonstrated the potential for a market segmentation from an external source to improve self-response propensity models using data from the 2010 Census and the American Community Survey.
- Demonstrated that market segmentation from an external source aid in providing useful information about problems in the Census enumeration of young children.
- Established theoretical limitations on consistent estimation of variance component parameters from informatively sampled complex survey data based only on single-inclusion weights.
- Developed a simple and novel measure of uncertainty for an estimated ranking with theory, using American Community Survey travel time to work data, and with a visualization.
- Extended the current *equal proportions* methodology by appealing to probability sampling results.
- Developed a general exact optimal sample allocation algorithm with bounded cost and bounded stratum sample sizes.

Short-Term Activities (FY 2021 – FY 2023):

- Extend Machine Learning approaches to non-response segmentation and frame changes.
- Develop optimal stratification in economic surveys, using multiple measures of size and multiple survey outcomes.
- Document biases of SDR design-based variance estimates for survey-weighted totals in small domains, and what survey design and attribute features they depend on.
- Continue research into post-stratified weight adjustment methodology and assessment of weights, with application to low-response probability surveys and non-probability data collection as in the Tracking Survey.
- Extend research into stratification methodology for economic surveys based on multiple MOS variables and multiple outcomes.
- Continue research into alternative techniques for statistical nondisclosure control motivated by randomize-response techniques
- Improve methodology for measuring uncertainty in rankings.
- Extend methodology for exact optimal sample allocation and apportionment.

Longer-Term Activities (beyond FY 2023):

- Extension of Census Matching capability to non-PIK persons using Administrative Records, Duplicate Status and Post-Enumeration Survey data for evaluation of Matching quality.
- Develop software that is re-usable and easily implementable for small area prediction within language minority groups in connection with the determinations of ballot language assistance by jurisdiction and American Indian Area under Section 203 of the Voting Rights Act.
- Further investigate the statistical implications and assumptions of formal privacy (e.g., differential privacy) methods in order to understand how the methods may impact the use of data products and to develop estimates of variability of released data that has been privatized by noise infusion.
- Develop statistical methods and theory related to the use of differential privacy to release data from unequal probability sampling surveys. A specific focus of this research would be on how to account for the sampling probabilities/weights in the planning of the privacy budget.
- Develop probability sampling methods targeted to the complement of an administrative records database within a survey frame such as the MAF; this research will require combining statistical models for joint dependence of administrative records and survey or census response, to be incorporated into new response propensity models in terms of which the survey data can be analyzed.
- Develop spatial models and associated small area estimation techniques in terms of Generalized Linear Mixed Models (GLMMs) with covariates recoded to incorporate local spatial geographic/demographic/economic effects, and compare the performance

of these models with Bayes-hierarchical models currently being developed elsewhere at the Census Bureau using American Community Survey data. Such GLMM spatial models may also be applicable to the evaluation of canvassing and address status changes in the MAF.

Selected Publications:

Mulry, M.H. and Mule, V.T. (2022). "Advances in the Use of Capture-Recapture Methodology in the Estimation of U.S. Census Coverage Error," In Recent Advances on Sampling Methods and Educational Statistics. In Honor of S. Lynne Stokes. Editors Hon Keung Tony Ng and Daniel F. Heitjan, 93–116, ISSN 2524-7735, https://doi.org/10.1007/978-3-031-14525-4

Slud, E., Hall, A., and Franco, C. (In Press). "Small Area Estimates for Voting Rights Act Section 203(b) Coverage Determinations," *Calcutta Statistical Association Bulletin*.

Nayak, T.K. (2021). "A Review of Rigorous Randomized Response Methods for Protecting Respondent's Privacy and Data Confidentiality," in *Methodology and Applications of Statistics: A Volume in Honor of C.R. Rao on the Occasion of his 100th Birthday*, ed. B.C. Arnold, N. Balakrishnan and C.A. Coelho, New York: Springer, pp. 319-341.

Wright, T. (2021). "From Cauchy-Schwartz to the House of Representatives: Application of Lagrange's Identity," *Mathematics Magazine, Vol 94,* 244-256.

Mulry, M., Bates, N., and Virgile, M. (2021). "Viewing Participation in Censuses and Surveys through the Lens of Lifestyle Segments" (print), *Journal of Survey Statistics and Methodology*, doi:1093/jssam/smaa006.

Zhai, X., and Nayak, T.K. (2021). "A Post-randomization Method for Rigorous Identification Risk Control in Releasing Microdata," *Journal of Statistical Theory and Practice,* 15, Article 8, https://doi.org/10.1007/s42519-020-00143-2.

Trudell, T., Dong, K., Slud, E., and Cheng, Y. (In Press). "Computing Replicated Variance for Stratified Systematic Sampling," *Proceedings of the Survey Research Methods Section of the American Statistical Association*.

Wright, T. (2020). "A General Exact Optimal Sample Allocation Algorithm: With Bounded Cost and Bounded Sample Sizes," *Statistics and Probability Letters, Vol 165,* Article 108829.

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

Franco, C., Little, R., Louis, T., and Slud, E. (2019). "Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys," *Journal of Survey Statistics and Methodology, 7,* 334-364.

Slud, E. and Thibaudeau, Y. (2019). "Multi-Outcome Longitudinal Small Area Estimation, A Case Study," *Statistical Theory and Related Fields. Special Issue on Small Area Estimation, 3,* 136-149.

Wright, T., Klein, M., and Wieczorek, J. (2019). "A Primer on Visualizations for Comparing Populations, Including the Issue of Overlapping Confidence Intervals," *The American Statistician, Vol 73, No 2,* 165-178.

Chai, J. and Nayak, T. (2018). "A Criterion for Privacy Protection in Data Collection and its Attainment via Randomized Response Procedures," Electronic Journal of Statistics 12 (2), 4264-4287.

de Oliveira, V., Wang, B., and Slud, E. (2018). "Spatial Modeling of Rainfall Accumulated over Short Periods of Time," *Journal of Multivariate Analysis,* 166, 129-149.

Dong, K., Trudell, T., Slud, E., and Cheng, Y. (2018). "Understanding Variance Estimator Bias in Stratified Two-Stage Sampling," *Proceedings of the Survey Research Methods Section of the American Statistical Association*.

Klein, M., Wright, T., and Wieczorek, J. (2018). "A Simple Joint Confidence Region for A Ranking of K Populations: Application to American Community Survey's Travel Time to Work Data," *Research Report Series (Statistics #2018-04),* Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.

Lu, B. and Ashmead, R. (2018). "Propensity Score Matching Analysis for Causal Effects with MNAR Covariates," *Statistica Sinica, 28,* 2005-2025.

Mulry, M.H, Kaputa, S., and Thompson, K. (2018). "Initial M-estimation Parameter Settings for Detection and Treatment of Influential Values," *Journal of Official Statistics, 34(2).* 483–501. http://dx.doi.org/10.2478/JOS-2018-0022

Nayak, T., Zhang, C., and You, J. (2018). "Measuring Identification Risk in Microdata Release and Its Control by Post-randomisation," *International Statistical Review, 86 (2)*, 300-321.

Slud, E., Vonta, I., and Kagan, A. (2018). "Combining Estimators of a Common Parameter across Samples," *Statistical Theory and Related Fields*, *2*, 158-171.

Wright, T. (2018). "No Calculation When Observation Can Be Made," in A.K. Chattopadhyay and G. Chattopadhyay (Eds), *Statistics and Its Applications*, Springer Singapore, 139-154.

Ashmead, R., Slud, E., and Hughes, T. (2017). "Adaptive Intervention Methodology for Reduction of Respondent Contact Burden in the American Community Survey," *Journal of Official Statistics, 33(4), 901-919.*

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.

Mulry, M.H. and Keller, A. (2017). "Comparison of 2010 Census Nonresponse Follow-up 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

Mulry, M.H., 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.

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.

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. (2017). “Exact Optimal Sample Allocation: More Efficient Than Neyman,” *Statistics and Probability Letters, 129,* 50-57.

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., 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

Nayak, T. and Adeshiyan, S. (2016). “On Invariant Post-randomization for Statistical Disclosure Control,” International Statistical Review, 84 (1), 26-42.

Nayak, T., Adeshiyan, S. and Zhang, C. (2016). “A Concise Theory of Randomized Response Techniques for Privacy and Confidentiality Protection,” Handbook of Statistics, 34, 273-286.

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, D.C.

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*.

Slud, Eric. (2015). “Impact of Mode-based Imputation on ACS Estimates,” *American Community Survey Research and Evaluation Memorandum,* #ACS-RER-O7.

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.

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.

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. 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.

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.

Tang, M., Slud, E., and Pfeiffer, R. (2014). “Goodness of Fit Tests for Linear Mixed Models,” *Journal of Multivariate Analysis, 130*, 176-193.

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, D.C.

Wright, T. (2014). “Lagrange’s Identity and Congressional Apportionment,” *The American Mathematical Monthly, 121*, 523-528.

Slud, E., Grieves, 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.

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, D.C.

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.

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, U.S. Census Bureau, Wash. D.C.

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.

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.

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, D.C.

Contact: Eric Slud, Mary Mulry, Michael Ikeda, Patrick Joyce, Tapan Nayak, Edward H. Porter, Tommy Wright

Funding Sources for FY 2021-2025: 0331 – Working Capital Fund / General Research Project
Various Decennial, Demographic, and Economic Projects

Time Series & Seasonal Adjustment

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

Research Problems:

- All contemporary seasonal adjustment programs of interest depend heavily on time series models for trading day and calendar effect estimation, for modeling abrupt changes in the trend, for providing required forecasts, and, in some cases, for the seasonal adjustment calculations. Better methods are needed for automatic model selection, for detection of inadequate models, and for assessing the uncertainty in modeling results due to model selection, outlier identification and non-normality. Also, new models are needed for complex holiday and calendar effects.
- Diagnostics of seasonality must address differing sampling frequencies (monthly versus quarterly) and multiple forms of seasonality (cycles of annual versus weekly period), and must distinguish between raw and seasonally adjusted data.
- Multivariate modeling can not only provide increased precision of seasonal adjustments, but can also assist with series that have a low signal content. Moreover, multivariate techniques expand the class of univariate models, allowing the modeling of seasonal heteroscedasticity. This motivates the need to develop a viable multivariate seasonal adjustment methodology that can handle modeling, fitting, and seasonal adjustment of a large number of series.
- Time series data are being measured at higher sampling rates or over geographical regions, requiring new seasonal adjustment methods for high frequency/space-time data.
- Many published time series arise from sample surveys, and are subject to sampling error. Methodology and algorithms are needed to incorporate sampling error components into the existing seasonal adjustment framework.

Current Subprojects:

- Seasonal Adjustment (McElroy/ADRM, Livsey, Pang, Roy)
- Time Series Analysis (McElroy/ADRM, Livsey, Pang, Roy, Trimbur)

Potential Applications

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

Accomplishments (October 2018-September 2020):

- Developed and implemented new algorithms for ragged edge missing value imputation, and ad hoc filtering of multivariate time series.
- Implemented and tested autoregressive diagnostics for seasonality.
- Refined a benchmarking method to remove seasonality from indirect seasonal adjustments.
- Added new models with stable parameterizations to multivariate time series software.
- Studied an EM approach to modeling multivariate time series.
- Studied outlier processes, allowing for a new approach to extreme-value adjustment of seasonal time series.
- Developed methods and formulas for quadratic filtering and forecasting of time series.

Short-Term Activities (FY 2021 – FY 2023):

- Continue developing diagnostics for seasonality by refining the AR diagnostic and examining forecast error and partial autocorrelation.
- Continue the study of weekly and daily 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.
- Develop nonlinear filtering and prediction methods based on autocumulants, with applications to seasonal adjustment in the presence of extremes.
- Develop improved automatic model identification methods.
- Develop extensions to maximum entropy extreme-value framework, allowing for more general types of outliers.
- Generate an R package for Ecce Signum, and disseminate X-13 R Story.
- 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.

- Study the impact of sampling error on seasonal adjustment.

Longer-Term Activities (beyond FY 2023):

- 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.
- Improve the speed and stability of likelihood optimization in X-13ARIMA-SEATS.
- 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 treating sampling error and stochastic trading day.
- 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:

McElroy, T., Roy, A., and Hore, G. (2023). “FLIP: a Utility Preserving Privacy Mechanism for Time Series,” *Journal of Machine Learning Research*, 24, 1-29.

McElroy, T. and Politis, D. (2023). “Estimating the Spectral Density at Frequencies Near Zero,” *Journal of the American Statistical Association*, published online.

McElroy, T., Ghosh, D., and Lahiri, S. (2023). “Quadratic Prediction of Time Series via Autocumulants,” *Sankhya A*, published online.

McElroy, T. and Jach, A. (2023). “Identification of the Differencing Operator of a Non-stationary Time Series via Testing for Zeroes in the Spectral Density*,” Computational Statistics and Data Analysis*, 177, 107580.

McElroy, T. and Trimbur, T. (2022). “Variable Targeting and Reduction in Large Vector Autoregressions with Applications to Workforce Indicators,” *Journal of Applied Statistics*, 50, 1515-1537.

McElroy, T. and Politis, D. (2022). “Optimal Linear Interpolation of Multiple Missing Values,” *Statistical Inference for Stochastic Processes*, 25, 471-483.

McElroy, T. (2022). “Casting Vector Time Series: Algorithms for Forecasting, Imputation, and Signal Extraction,” *Electronic Journal of Statistics*, 16, 5534-5569.

McElroy, T. (2022). “Stationary Parameterization of GARCH Processes,” *Economics Bulletin*, 42 (4).

Davis, R.A., Fokianos, K., Holan, S., Joe, H., Livsey, J., Lund, R.B., Pipiras, V., and Ravishanker, N. (In Press). “Count Time Series: A Methodological Review,” *Journal of the American Statistical Association*.

Binder, C., McElroy, T., and Sheng, X. (2022). “Term Structure of Uncertainty: New Evidence from Survey Expectations,” *Journal of Money, Credit, and Banking, 54(1),* 39-71.

Chen, B., McElroy, T., and Pang, O. (2022). “Assessing Residual Seasonality in the U.S. National Income and Product Accounts Aggregates,” *Journal of Official Statistics, Volume 38, Issue 2*, 399-428.

McElroy, T. (2022). “Frequency Domain Calculation of Seasonal VARMA Autocovariances,” *Journal of Computational and Graphical Statistics, 31(1),* 301-303.

McElroy, T. and Politis, D. (2022). “Optimal Linear Interpolation of Multiple Missing Values,” *Statistical Inference for Stochastic Processes*, 1-13.

McElroy, T. and Roy, A. (2022). “A Review of Seasonal Adjustment Diagnostics,” *International Statistical Review, 90(2),* 259-284.

McElroy, T. and Roy, A. (2022). “Model Identification via Total Frobenius Norm of Multivariate Spectra,” *Journal of the Royal Statistical Society, Series B, Volume 84*, 473-495.

McElroy, T. and Trimbur, T. (2022). “Variable Targeting and Reduction in Large Vector Autoregressions with Applications to Workforce Indicators,” *Journal of Applied Statistics*, 1-23.

Trimbur, T. and McElroy, T. (2022). “Modelled Approximations to the Ideal Filter with Application to GDP and Its Components,” *The Annals of Applied Statistics, 16(2),* 627-651.

Jia, Y., Kechagias, S., Livsey, J., Lund, R., Pipiras, V. (2021). “Latent Gaussian Count Time Series Modelling,” *Journal of the American Statistical Association*.

McElroy, T. (2021). “A Diagnostic for Seasonality Based upon Polynomial Roots of ARMA Models,” *Journal of Official Statistics, 37(2)*, 1-28.

McElroy, T. and Das, S. (2021). “Nonlinear Prediction via Hermite Transformation,” *Statistical Theory and Related Fields 5(1),* 49-54.

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

McElroy, T., Roy, A., Livsey, J., Firestine, T., and Notis, K. (2021). “Anticipating Revisions in the Transportation Services Index,” *Journal of the International Association of Official Statistics*, *37*, 641-653.

Baker, S., McElroy, T.S., and Sheng, X. (2020). “Expectation Formation Following Large and Unpredictable Shocks,” *Review of Economics and Statistics, 14,* 112-130.

McElroy, T.S. and Politis, D.N. (2020). *Time Series: a First Course with Bootstrap Starter*. New York: Chapman Hall.

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

Hyatt, H. and McElroy, T.S. (2019). “Labor Reallocation, Employment, and Earnings: Vector Autoregression Evidence,” *LABOUR, 33(4),* 463-487.

McElroy, T.S. and Jach, A. (2019). “Testing Collinearity of Vector Time Series,” *The Econometrics Journal, 22(2),* 97-116.

McElroy, T. S. and Jach, A. (2019). “Subsampling Inference for the Autocorrelations of GARCH Processes,” Published online, *Journal of Financial Econometrics, 17(3),* 495-515.

McElroy, T. S., Pang, O., and Sheldon, G. (2019). “Custom Epoch Estimation for Surveys,” Published online, *Journal of Applied Statistics, 46,* 638-663.

McElroy, T.S. and Penny, R. (2019). “Maximum Entropy Extreme-Value Seasonal Adjustment,” *Australian New Zealand Journal of Statistics, 61(2),* 152-174.

Roy, A., McElroy, T. S., and Linton, P. (2019). “Estimation of Causal Invertible VARMA Models,” *Statistica Sinica, 29(1),* 455-478.

Wildi, M. and McElroy, T. S. (2019). “The Trilemma between Accuracy, Timeliness, and Smoothness in Real-Time Signal Extraction,” *International Journal of Forecasting, 35,* 1072-1084

Findley, D.F. and McElroy, T. S. (2018). “Background and Perspectives for ARIMA Model-Based Seasonal Adjustment,” *Research Report Series (Statistics #2018-07),* Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.

Lin, W., Huang, J., and McElroy, T. S. (2018). “Time Series Seasonal Adjustment Using Regularized Singular Value Decomposition,” Published online, *Journal of Business and Economics Statistics.*

Livsey, J., Lund, R., Kechagias, S., and Pipiras, V. (2018). “Multivariate Integer-valued Time Series with Flexible Autocovariances and Their Application to Major Hurricane Counts,” *Annals of Applied Statistics, 12(1):* 408-431.

McElroy, T. S. (2018). “Recursive Computation for Block Nested Covariance Matrices,” *Journal of Time Series Analysis, 39 (3),* 299-312.

McElroy, T. S. (2018). “Seasonal Adjustment Subject to Accounting Constraints,” *Statistica Neerlandica, 72,* 574-589.

McElroy, T.S., Monsell B. C., and Hutchinson, R. (2018). “Modeling of Holiday Effects and Seasonality in Daily Time Series,” *Research Report Series (Statistics 2018-01),* Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.

McElroy, T.S. and Roy, A. (2018). “Model Identification via Total Frobenius Norm of Multivariate Spectra,” *Research Report Series (Statistics #2018-03),* Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.

McElroy, T. S. and Roy, A. (2018). “The Inverse Kullback Leibler Method for Fitting Vector Moving Averages,” *Journal of Time Series Analysis, 39,* 172-191.

Nagaraja, C. and McElroy, T. S. (2018). “The Multivariate Bullwhip Effect,” *European Journal of Operations Research, 267,* 96-106.

Blakely, C. and McElroy, T. S. (2017). “Signal Extraction Goodness-of-fit Diagnostic Tests under Model Parameter Uncertainty: Formulations and Empirical Evaluation,” *Econometric Reviews, 36 (4),* 447-467.

Findley, D.F., Lytras, D. P., and McElroy, T. S. (2017). “Detecting Seasonality in Seasonally Adjusted Monthly Time Series,” *Research Report Series (Statistics #2017-03),* Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.

Holan, S., McElroy, T. S., and Wu, G. (2017). “The Cepstral Model for Multivariate Time Series: The Vector Exponential Model,” *Statistica Sinica 27,* 23-42.

McElroy, T. S. (2017). “Computation of Vector ARMA Autocovariances,” *Statistics and Probability Letters, 124,* 92-96.

McElroy, T. S. (2017). “Multivariate Seasonal Adjustment, Economic Identities, and Seasonal Taxonomy,” *Journal of Business and Economics Statistics, 35 (4),* 511-525.

McElroy, T. S. and McCracken, M. (2017). “Multi-Step Ahead Forecasting of Vector Time Series,” *Econometric Reviews, 36 (5),* 495-513.

McElroy, T.S. and Monsell, B. C (2017). “Issues Related to the Modeling and Adjustment of High Frequency Time Series,” *Research Report Series (Statistics #2017-08),* Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.

Sanyal, A., Mitra, P., McElroy, T.S., and Roy, A. (2017). “Holiday Effects in Indian Manufacturing Series,” *Research Report Series (Statistics #2017-04),* Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.

Trimbur, T. and McElroy, T. S. (2017). “Signal Extraction for Nonstationary Time Series with Diverse Sampling Rules,” *Journal of Time Series Econometrics, 9 (1).*

Janicki, R. and McElroy, T. (2016). “Hermite Expansion and Estimation of Monotonic Transformations of Gaussian Data,” *Journal of Nonparametric Statistics, 28(1),* 207-234.

McElroy, T. S. (2016). “Non-nested Model Comparisons for Time Series,” *Biometrika, 103,* 905-914.

McElroy, T. (2016). “On the Measurement and Treatment of Extremes in Time Series,” *Extremes, 19(3),* 467-490.

McElroy, T. and Nagaraja, C. (2016). “Tail Index Estimation with a Fixed Tuning Parameter Fraction,” *Journal of Statistical Planning and Inference*, *170*, 27-45.

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.

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. (2015). “When are Direct Multi-Step and Iterative Forecasts Identical?” *Journal of Forecasting, 34***,** 315-336.

McElroy, T. and Findley, D. (2015). “Fitting Constrained Vector Autoregression Models,” in *Empirical Economic and Financial Research*.

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 Pang, O. (2015). “The Algebraic Structure of Transformed Time Series,” in *Empirical Economic and Financial Research*.

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 Holan, S. (2014). “Asymptotic Theory of Cepstral Random Fields,” *Annals of Statistics*, 42, 64-86.

McElroy, T. and Maravall, A. (2014). “Optimal Signal Extraction with Correlated Components,” *Journal of Time Series Econometrics*, *6*, 237--273.

McElroy, T. and Monsell, B. (2014). “The Multiple Testing Problem for Box-Pierce Statistics,” *Electronic Journal of Statistics*, *8*, 497-522.

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.

Monsell, B. C. (2014) “The Effect of Forecasting on X-11 Adjustment Filters,” *2014 Proceedings American Statistical Association* [CD-ROM]: Alexandria, VA.

Roy, A., McElroy, T., and Linton, P. (2014). “Estimation of Causal Invertible VARMA Models,” *Cornell University Library*, http://arxiv.org/pdf/1406.4584.pdf.

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, D.C.

McElroy, T. (2013). “Forecasting CARIMA Processes with Applications to Signal Extraction,” *Annals of the Institute of Statistical Mathematics*, *65*, 439-456.

McElroy, T. and Politis, D. (2013). “Distribution Theory for the Studentized Mean for Long, Short, and Negative Memory Time Series,” *Journal of Econometrics*, *177*, 60-74.

McElroy, T. and Wildi, M. (2013). “Multi-Step Ahead Estimation of Time Series Models,” *International Journal of Forecasting 29*, 378-394.

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.*

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.*

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.

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 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 Politis, D. (2012). "Fixed-b Asymptotics for the Studentized Mean for Long and Negative Memory Time Series," *Econometric Theory, 28*, 471-481.

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.

Contact: Tucker McElroy (R&M), James Livsey, Osbert Pang, Anindya Roy, Bill Bell (R&M).

Funding Sources for FY 2021-2025: 0331 – Working Capital Fund / General Research Project
Economic Projects

Experimentation, Prediction, & Modeling

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

Research Problems:

- Investigate established methods and novel extensions to support design (e.g., factorial designs), analysis, and sample size determination for Census Bureau experiments.
- Investigate methodology for experimental designs embedded in sample surveys, including large-scale field experiments embedded in ongoing surveys. This includes design-based and model-based analysis and variance estimation incorporating the sampling design and the experimental design (van den Brakel, Survey Methodology, 2005); factorial designs embedded in sample surveys (van den Brakel, Survey Methodology, 2013), and the estimation of interactions; and testing non-response using embedded experiments.
- Identify and develop statistical models (e.g., loglinear models, mixture models, and mixed-effects models), associated methodologies, and computational tools for problems relevant to the Census Bureau.
- Assess the applicability of *post hoc* methods (e.g., multiple comparisons and tolerance intervals) with future designed experiments and when reviewing previous data analyses.
- 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.
- Develop a technique for mis-reporting via the COM-Poisson distribution in order to estimate true counts.
- Develop a disclosure policy motivated by the COM-Poisson and related distributions that allows one to protect individual information reported in two-way and multi-way tables.

Current Subprojects:

- Developing Flexible Distributions and Statistical Modeling for Count Data Containing Dispersion (Sellers, Morris, Raim).
- Design and Analysis Methods for Experiments (Raim, Mathew, Sellers)

Potential Applications:

- Modeling can help to characterize relationships between variables measured in censuses, sample surveys, and administrative records and quantify their uncertainty.
- 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 censuses and surveys. Sample sizes can be determined to achieve desired power using planned designs and statistical procedures.
- Embedded experiments can be used to evaluate the effectiveness of alternative contact strategies.
- The collection of experimental design procedures currently utilized with the American Community Survey can be expanded.
- 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. The importance of ratio edits and multivariate/multiple edits is noted in the work of Thompson and Sigman (*Journal of Official Statistics*, 1999) and de Waal, Pannekoek and Scholtus (Handbook of Statistical Data Editing and Imputation, 2011).
- Principled measures of statistical variability can be provided for constructs like the POP Division's Population Estimates.
- Mis-reporting techniques could be used to assess the amount of misreporting in historical Census datasets to aid in model development to estimate true survey count outcomes.
- Statistical disclosure limitation constructs would allow the Census Bureau to release statistical measures associated with a general distributional form while protecting individual privacy. These measures would allow one to estimate the form of multi-way tables of interest while masking the true outcomes.

Accomplishments (October 2018-September 2020):

- Completed paper on spatio-temporal change of support modeling in R and released stcos R package.
- Addressed issues with COM-Poisson normalizing constant in the COMPoissonReg R package.
- Completed paper on Conway-Maxwell (COM) multinomial distribution and its use in analyzing clustered multinomial datasets that exhibit over- or under-dispersion.
- Developed and released COMMultReg R package to support COM-multinomial paper.
- Completed paper on continuation-ratio logit modeling for sample size determination and analysis of experiments involving sequences of success/failure trials. Such models support the study of nonresponse probabilities under multiple enumeration attempts to each household.
- Completed paper on comparing pairs of discrete distributions via multinomial outcomes to determine if one is closer to a discrete uniform distribution. This was applied to Census Bureau call volume data to determine if a staggered mailing strategy leads to significantly more uniform call distributions than a simpler strategy where mail is sent to all recipients at once.
- Completed development of a one-step autoregressive model for count data motivated by the COM-Poisson distribution.

Short-Term Activities (FY 2021 – FY 2023):

- Explore COM-multinomial as a model for missing observations in clustered data under a Bayesian setting.
- Extend work on sample size determination with continuation-ratio logit model to a mixed effects setting.
- Develop a multivariate COM-Poisson distribution model.

Longer-Term Activities (beyond FY 2023):

- Develop generalized/flexible spatial and time series models motivated by the COM-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), Mukerjee, Dasgupta and Rubin (*Journal of the American Statistical Association*, 2018), Branson and Dasgupta (*International Statistical Review*, 2020). 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).
- Generalize the Kadane et al. (2006) COM-Poisson motivated data disclosure limitation procedure for one-way tables to handle two-way and multi-way tables. Determine the associated sufficient statistics of the bivariate (or multivariate) COM-Poisson distribution and use them to describe the space of feasible tables that can be used to substitute the true contingency table.
- Consider generalizations of the frequentist and Bayesian approaches to address under-reporting described in Winkelmann (1996), Fader and Hardie (2000), Neubauer and Djuras (2009), and Neubauer et al. (2009) to allow for data dispersion via the COM-Poisson distribution.

Selected Publications:

Raim, A.M., Nichols, E., and Mathew, T. (2023). "A Statistical Comparison of Call Volume Uniformity Due to Mailing Strategy," *Journal of Official Statistics*, 39, 103-121.

Raim, A.M., Mathew, T., Sellers, K. F., Ellis, R., and Meyers, M. (2023). "Design and Sample Size Determination for Experiments on Nonresponse Follow-up using a Sequential Regression Model," *Journal of Official Statistics*, 39(2), 173-202.

Raim, A.M. (2023). "Direct Sampling with a Step Function," *Statistics and Computing*, 33(22). https://doi.org/10.1007/s11222-022-10188.

Lucagbo, M., Mathew, T., and Young, D. (2023). "Rectangular Multivariate Normal Prediction Regions for Setting Reference Regions in Laboratory Medicine," *Journal of Biopharmaceutical Statistics*, 33(2), 191-209.

Lucagbo, M. and Mathew, T. (2023). "Rectangular Tolerance Regions and Multivariate Normal Reference Regions in Laboratory Medicine," *Biometrical Journal*, 65(3).

Arsham, A., Bebu, I., and Mathew, T. (2023). "Cost-Effectiveness Analysis Under Multiple Effectiveness Outcomes: A Probabilistic Approach," *Statistics in Medicine*, 42, 3936-3955.

Arsham, A., Bebu, I., and Mathew, T. (2022). "A Bivariate Regression-Based Cost-Effectiveness Analysis," *Journal of Statistical Theory and Practice, 16,* Article No. 27.

Janicki, R., Raim, A.M., Holan, S.H., and Maples, J. (2022). "Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models with Application to American Community Survey Special Tabulations," *The Annals of Applied Statistics, Volume 16, Issue 1*, 144-168.

Lucagbo, M. and Mathew, T. (2022). "Rectangular Confidence Regions and Prediction Regions in Multivariate Calibration," *Journal of the Indian Society for Probability and Statistics*, *23*, 155–171.

Morris, D.S. and Sellers, K.F. (2022). "A Flexible Mixed Model for Clustered Count Data," *Stats: Special Issue on Statistics, Data Analytics, and Inferences for Discrete Data, 5(1):* 52–69. https://doi.org/10.3390/stats5010004.

Rivas, A., Antoun, C., Feuer, S., Mathew, T., Nichols, E., Olmsted-Hawala, E. and Wang, L (2022), "Comparison of Three Navigation Button Designs in Mobile Survey for Older Adults," *Survey Practice, 15(1).*

Weems, K.S., Sellers, K.F., and Li, T. (2021). "A Flexible Bivariate Distribution for Count Data Expressing Data Dispersion,"

Communications in Statistics - Theory and Methods, https://doi.org/10.1080/03610926.2021.1999474.

Feng, X., Mathew, T, and Adragni, K. (2021). "Interval Estimation of the Intra-class Correlation in General Linear Mixed Effects Models," *Journal of Statistical Theory and Practice*, *15, Article 65*.

Sellers, K.F., Arab, A., Melville, S., and Cui, F. (2021). "A Flexible Univariate Moving Average Time-Series Model for Dispersed Count Data," *Journal of Statistical Distributions and Applications 8 (1).* https://doi.org/10.1186/s40488-021-00115-2

Sellers, K.F., Li, T., Wu, Y., and Balakrishnan, N. (2021). "A Flexible Multivariate Distribution for Correlated Count Data," *Stats, 4(2),* 308-326, https://doi.org/10.3390/stats4020021.

Zhao, J., Mathew, T., and Bebu, I. (2021). "Accurate Confidence Intervals for Inter-Laboratory Calibration and Common Mean Estimation," *Chemometrics and Intelligent Laboratory Systems*, *208*. DOI: 10.1016/j.chemolab.2020.104218.

Zimmer, Z., Park, D., and Mathew, T. (2021). "Tolerance Limits under Zero-Inflated Lognormal and Gamma Distributions," *Computational and Mathematical Methods*, *Special Issue on Statistics, 3*. DOI: 10.1002/cmm4.1113.

Morris, D.S., Raim, A.M., and Sellers, K.F. (In Press). "A Conway-Maxwell-multinomial Distribution for Flexible Modeling of Clustered Categorical Data," *Journal of Multivariate Analysis*. DOI: https://doi.org/10.1016/j.jmva.2020.104651.

Raim, A.M., Holan, S.H., Bradley, J.R., and Wikle, C.K. (2020). stcos: "Space-Time Change of Support, version 0.3.0," https://cran.r-project.org/package=stcos.

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

Zhu, L., Sellers, K., Morris, D., Shmueli, G., and Davenport, D. (2020). cmpprocess: "Flexible Modeling of Count Processes," version 1.1, https://cran.r-project.org/package=cmpprocess

Raim, A.M., Holan, S.H., Bradley, J.R., and Wikle, C.K. (2019). "Spatio-Temporal Change of Support Modeling for the American Community Survey with R," URL: https://arxiv.org/abs/1904.12092.

Sellers, K., Lotze, T., and Raim, A. (2019). COMPoissonReg: "Conway-Maxwell-Poisson Regression, version 0.7.0," https://cran.r-project.org/package=COMPoissonReg

Sellers, K.F. and Young, D. (2019). "Zero-inflated Sum of Conway-Maxwell-Poissons (ZISCMP) Regression with Application to Shark Distributions," *Journal of Statistical Computation and Simulation*, *89 (9):* 1649-1673.

Sellers, K., Morris, D., Balakrishnan, N., and Davenport, D. (2018). multicmp: "Flexible Modeling of Multivariate Count Data via the Multivariate Conway-Maxwell-Poisson Distribution," version 1.1, https://cran.r-project.org/package=multicmp

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., Holan, S.H., Bradley, J.R., and Wikle, C.K. (2017). "A Model Selection Study for Spatio-Temporal Change of Support," in *Proceedings, Government Statistics Section of the American Statistical Association,* Alexandria, VA: American Statistical Association.

Sellers, K.F., and Morris, D. (2017). "Under-dispersion Models: Models That Are 'Under The Radar'," *Communications in Statistics – Theory and Methods*, *46 (24):* 12075-12086.

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*.

Young, D.S., Raim, A.M., and Johnson, N.R. (2017). "Zero-inflated Modelling for Characterizing Coverage Errors of Extracts from the U.S. 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.

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.

Mathew, T., Menon, S., Perevozskaya, I., and Weerahandi, S. (2016). "Improved Prediction Intervals in Heteroscedastic Mixed-Effects Models," *Statistics & Probability Letters, 114*, 48-53.

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.F., Morris, D.S., and Balakrishnan, N. (2016). "Bivariate Conway-Maxwell-Poisson Distribution: Formulation, Properties, and Inference," *Journal of Multivariate Analysis, 150:*152-168.

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.

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.

Young, D. and Mathew, T. (2015). "Ratio Edits Based on Statistical Tolerance Intervals," *Journal of Official Statistics 31*, 77-100.

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

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

Gamage, G., Mathew, T., and Weerahandi, S. (2013). “Generalized Prediction Intervals for BLUPs in Mixed Models,” *Journal of Multivariate Analysis*, *120*, 226-233.

Mathew, T. and Young, D. S. (2013). “Fiducial-Based Tolerance Intervals for Some Discrete Distributions,” *Computational Statistics and Data Analysis*, 61, 38-49.

Young, D.S. (2013). “Regression Tolerance Intervals,” *Communications in Statistics – Simulation and Computation*, 42(9), 2040-2055.

Contact: Andrew Raim, Thomas Mathew, Kimberly Sellers, Darcy Morris

Funding Sources for FY 2021-2025: 0331 – Working Capital Fund / General Research Project
Various Decennial and Demographic Projects

Simulation, Data Science, & Visualization

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

Research Problems:
- Systematically develop an environment for simulating complex sample surveys that can be used as a test-bed for new data analysis methods.
- Develop new methods for statistical disclosure control that simultaneously protect confidential data from disclosure while enabling valid inferences to be drawn on relevant population parameters.
- Develop models for the analysis of measurement errors in Demographic sample surveys (e.g., Current Population Survey or the Survey of Income and Program Participation).
- Investigate noise infusion and synthetic data for statistical disclosure control.

Current Subprojects:
- Development and Evaluation of Methodology for Statistical Disclosure Control (Nayak)
- The Ranking Project: Methodology Development and Evaluation (Wright,Klein/FDA,Wieczorek/Colby College, Yau)

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

Accomplishments (October 2018-September 2020):
- Developed and published visualizations for comparing populations.
- Developed and published theory and a visualization for expressing uncertainty and an overall ranking of populations.
- Conducted research and published results on randomized response methods for protecting respondent’s privacy and data confidentiality.

Short-Term Activities (FY 2021 – FY 2023):
- Continue development of new methodology for statistical disclosure control and to evaluate properties of new and existing methods.
- Improve visualizations for comparing populations and for overall rankings of populations.

Longer-Term Activities (beyond FY 2023):
- Study ways of quantifying the privacy protection/data utility tradeoff in statistical disclosure control.
- Create an environment for simulating complex aspects of economic/demographic sample surveys.
- Develop methodology for quantifying uncertainty in statistical rankings, and refine visualizations.

Selected Publications:

Guin, A., Roy, A., and Sinha, B. (2022). “Bayesian Analysis of Multiply Imputed Synthetic Data under the Multiple Linear Regression Model,” *International Journal of Statistical Sciences*, Volume 22(2), 25-38.

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

Moura, R., Klein, M., Zylstra, J., Coelho, C., and Sinha, B. (In Press). “Inference for Multivariate Regression Model Based on Synthetic Data Generated Under Plug-In Sampling,” *Journal of the American Statistical Association (Theory & Methods)*.

Chai, J. and Nayak, T.K. (2021). “Minimax Randomized Response Methods for Protecting Respondent’s Privacy,” *Communications in Statistics - Theory and Methods*, https://doi.org/10.1080/03610926.2021.1973503

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

Klein, M.D., Zylstra, J., and Sinha, B.K. (2019). “Finite Sample Inference for Multiply Imputed Synthetic Data under a Multiple Linear Regression Model,” *Calcutta Statistical Association Bulletin*. https://doi.org/10.1177/0008068318803814

Wright, T., Klein, M., and Wieczorek, J. (2019). “A Primer on Visualizations for Comparing Populations Including the Issue of Overlapping Confidence Intervals,” *The American Statistician*, Vol 73, No. 2, 165-178.

Chai, J. and Nayak, T.K. (2018). “A Criterion for Privacy Protection in Data Collection and Its Attainment via Randomized Response Procedures,” *Electronic Journal of Statistics*, *12*, 4264-4287.

Klein, M. and Datta, G. (2018). “Statistical Disclosure Control via Sufficiency under the Multiple Linear Regression Model,” *Journal of Statistical Theory and Practice, 12(1),* 100-110.

Nayak, T.K., Zhang, C., and You, J. (2018). “Measuring Identification Risk in Microdata Release and Its Control by Post-randomisation,” *International Statistical Review, 86(2)*, 300-321.

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 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/rrs2014-12.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.

Nayak, T.K., Sinha, B., and Zayatz, L. (2011). “Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection,” *Journal of Official Statistics, 27 (3)*, 527-544.

Sinha, B., Nayak, T.K., and Zayatz, L. (2011). “Privacy Protection and Quantile Estimation from Noise Multiplied Data,” *Sankhya*, Ser. B, 73, 297-315.

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: Tommy Wright, Tapan Nayak, Bimal Sinha, Nathan Yau

Funding Sources for FY 2021-2025: 0331 – Working Capital Fund / General Research Project
Various Decennial, Demographic, and Economic Projects

Cross-Cutting Statistical General Research Priorities

(Study/Working Group Members)

A. Design and Analysis of Sample Surveys around Administrative or Commercial Observational Databases, or based on Web (Opt-in) Data-Collections: specific analysis and focus could be directed at the Census Bureau Contact Frame (MAF subset with telephone and email addresses) as used e.g., in the current Household Pulse Survey.

LEADER: Eric Slud
(FIRST Choice: Darcy, Chad, Eric, Michael I., Mary, Soumen, Dan, Tommy, Tapan; SECOND Choice: Anindya)

This initiative involves new methodology including modeling, to support Census Bureau efforts to design sample surveys and censuses in the future around special national lists as frames. Such lists may be convenient because of auxiliary data they contain, such as administrative records, or because they refer to address lists with auxiliary validated contact information such as telephone numbers or email/IP addresses. The CSRM effort includes descriptive statistical summaries of the predictive characteristics of membership on one or more lists of these types, leading to the development of effective predictive models to be used in future designs in tandem with the general MAF frame.

B. Optimization-based (single-stage) approaches to Weight-adjustment for Probability and Nonprobability Samples.

LEADER: Emanuel Ben-David
(FIRST Choice: Emanuel; SECOND Choice: Mary, Isaac; OTHER Choice: Eric, Bimal)

Several of the Census Bureau's most important household surveys produce survey weights after several (up to 15!) successive difficult-to-document stages of adjustment or poststratification, with the result that the adjustments made in early stages are somewhat distorted in later stages. Methodology exists to do such poststratification in the form of Generalized Raking or Calibration by an optimization approach to minimize the degree of adjustment of base weights while ensuring exact or approximate conformity with calibration constraints to adhere to Population Estimates or other external-source totals for key variables. Considering the important application of Census Bureau surveys by survey methodologists to calibrate their own surveys for other purposes, this optimization-based approach would at the same time be easier to document and would maintain better simultaneous conformity with population controls than current methods.

C. Research on model-based imputation incorporating (nonrandom) Hot-deck values as covariates, leveraging descriptive analyses of differences between the hot-deck donor universe and general population.

LEADER:
(FIRST Choice: Bimal; SECOND Choice: Darcy, Eric, Soumen, Anindya, Jun; OTHER Choice: Kim)

The Census Bureau relies throughout its household surveys on whole-unit and single-item imputation methodology based on hot-deck algorithms to impute or allocate missing data from data supplied nearby responding units. Attempts to update these methods with model-based improvements have generally failed, at least in part because demographic predictive variables omit important neighborhood information that is obtained from nearby donor units. Taking such donor information into account within predictive models, instead of using it directly in imputation, is an approach that has not been adequately explored, and that the Census Bureau is uniquely situated to implement properly. Research along these lines would clarify the differences between single and joint distributions of donor versus general-population household variables, and could improve the assessment and representativeness of joint distributions of variables in microdata, which are always partially imputed.

D. Development of Model Diagnostics and Cross-validation methods for Imputation and Small Area Estimation models.

LEADER:
(FIRST Choice: Ryan, Gauri, Isaac, Kyle; SECOND Choice: Jerry, Kim; OTHER Choice: Maria, Soumen, Darcy)

Throughout Census Bureau research efforts, model-based methods for response propensity prediction, for unit and item imputation, and for small-area estimates are impeded by the lack of systematic methodology for assessment involving ground truth. Ad hoc model diagnostics generally reveal only the differences between the results from competing models. Methods of cross-validation – currently under-developed in small-area and survey-sampling literature, would improve the Census Bureau's ability to ensure quality of released data, supported by increased use of post-enumeration survey results from the decennial census.

E. Development of Survey and Sampling Microsimulation utility for applications to Nondisclosure (Synthetic Data), and to Test-beds for Model- and Design-based methods in Variance Estimation and (area- and unit-level) Small Area Estimation.

LEADER: Jerry Maples
(FIRST Choice: Jerry; SECOND Choice: Isaac, Osbert, Joe, Tapan: OTHER Choice: Soumen, Gauri)

A system for microsimulation of artificial-population survey and census data, in the context both of household and economic surveys, would have at least two important ongoing applications: (i) as test-beds for current and new model-based methods for producing imputations and special-purpose and small area data, and (ii) in the development of new methods for the release of partially synthetic data products whose nondisclosure properties and variability can be documented scientifically. Work along this line is already underway for some Economic surveys and for SAIPE testing.

F. Joint Time-Series/Spatial and Sampling Estimation Models and Diagnostics.

LEADER:
(FIRST Choice: Anindya, Kim, Osbert, Patrick, Jim; SECOND Choice: Ryan, Gauri, Joe, Dan; OTHER Choice: Soumen)

Time series expertise could be leveraged toward providing specialized data tabulations, customized to special-purpose time periods and small areas, if research were expanded on models and estimation methods jointly incorporating time series and sampling errors. There is considerable expertise in CSRM on time series forecasting, on demographic and on small area modeling and benchmarking. Development of time series methods for custom tabulations would promise new custom data products as well as new tools for assessing small-area estimates produced throughout the Census Bureau.

G. Other items (?) e.g., Methods of Assessing Variability of Census or Survey Totals based on Noise-Infused Data; random-based causal inference for complex experiments; design and analysis of experiments on non-response using sequential regression models; entity resolution, visualizations; small area estimation/longitudinal studies. (FIRST Choice: Thomas, Andrew, Beka, Jim, Nathan; SECOND Choice: Emanuel, Kyle, Chad, Michael I., Bimal, Tommy)