Effects of a Government-Academic Partnership: Has the NSF-Census Bureau Research Network Helped Secure the Future of the Federal Statistical System?

by

Daniel H. Weinberg, DHW Consulting and U.S. Census Bureau (retired)
John M. Abowd, Cornell University and U.S. Census Bureau
Robert F. Belli, University of Nebraska
Noel Cressie, University of Wollongong and University of Missouri
David C. Folch, Florida State University
Scott H. Holan, University of Missouri and U.S. Census Bureau
Margaret C. Levenstein, University of Michigan
Kristen M. Olson, University of Nebraska and U.S. Census Bureau
Jerome P. Reiter, Duke University and U.S. Census Bureau
Matthew D. Shapiro, University of Michigan
Jolene Smyth, University of Nebraska
Leen-Kiat Soh, University of Nebraska
Bruce D. Spencer, Northwestern University
Seth E. Spielman, University of Colorado
Lars Vilhuber, Cornell University and U.S. Census Bureau
Christopher K. Wikle, University of Missouri

11 October 2017

ABSTRACT

The National Science Foundation-Census Bureau Research Network (NCRN) was established in 2011 to create interdisciplinary research nodes on methodological questions of interest and significance to the broader research community and to the Federal Statistical System (FSS), particularly the Census Bureau. The activities to date have covered both fundamental and applied statistical research and have focused at least in part on the training of current and future generations of researchers in skills of relevance to surveys and alternative measurement of economic units, households, and persons. This paper discusses some of the key research findings of the eight nodes, organized into six topics: (1) Improving census and survey data collection methods; (2) Using alternative sources of data; (3) Protecting privacy and confidentiality by improving disclosure avoidance; (4) Using spatial and spatio-temporal statistical modeling to improve estimates; (5) Assessing data cost and quality tradeoffs; and (6) Combining information from multiple sources. It also reports on collaborations across nodes and with federal agencies, new software developed, and educational activities and outcomes. The paper concludes with an evaluation of the ability of the FSS to apply the NCRN's research outcomes and suggests some next steps, as well as the implications of this research-network model for future federal government renewal initiatives.

ACKNOWLEDGEMENTS

This paper began as a May 8, 2015 presentation to the National Academies of Science's Committee on National Statistics by two of the principal investigators of the National Science Foundation-Census Bureau Research Network (NCRN) – John Abowd and the late Steve Fienberg (Carnegie Mellon University). The authors acknowledge the contributions of the other principal investigators of the NCRN who are not co-authors of the paper (William Block, William Eddy, Alan Karr, Charles Manski, Nicholas Nagle, and Rebecca Nugent), the co-principal investigators, and the comments of Patrick Cantwell, Constance Citro, Adam Eck, Brian Harris-Kojetin, and Eloise Parker. We note with sorrow the deaths of Stephen Fienberg and Allan McCutcheon, two of the original NCRN principal investigators. The conclusions reached in this paper are not the responsibility of the National Science Foundation (NSF), the Census Bureau, or any of the institutions to which the authors belong. The NCRN was supported by NSF grants to the participating institutions: 1129475 to Northwestern University; 1130706 to Carnegie Mellon University; 1131500 to University of Michigan-Ann Arbor; 1131848 to Cornell University; 1131897 to Duke University and National Institute of Statistical Sciences (NISS); 1132008 to University of Colorado-Boulder; 1132015 to University of Nebraska-Lincoln; 1132031 to University of Missouri; and 1507241 for the Coordinating Office (Cornell, Duke, and NISS). The principal investigators also wish to acknowledge Cheryl Eavey's sterling grant administration on behalf of the NSF.

Effects of a Government-Academic Partnership: Has the NSF-Census Bureau Research Network Helped Secure the Future of the Federal Statistical System?

I. INTRODUCTION

The birth of modern sample surveys came as a result of the Great Depression of the 1930s and the government's need to measure and understand the nature of unemployment. A wide range of factors, including the growing volume of survey requests from government, business, pollsters, and non-profit organizations, increasingly intrusive direct marketing, changing concerns about privacy, and shifting attitudes about the role of government in society has led to a decline in participation in household and business surveys over the past 25 years (see Atrostic et al. 2001; Nancarrow et al. 2004; Brick and Williams 2013; Tourangeau and Plewes 2013; Groves and Harris-Kojetin 2017). The decline in participation in government surveys lowers the quality and increases the cost of official statistics.

Coincidently, the "Information Age" has ushered in a transformation in the way people and businesses think about data and statistical information. Large-scale data and computationally intensive methods, popularly known as "big data," are laying the foundation for a paradigm shift in the way users and producers conceptualize statistical information. The U.S. Census Bureau, and implicitly all of its federal statistical agency peers, cannot ignore the consequences of this shift.[1] By 2020, more than 64% of the projected population of the U.S. will have been born after the first email message was sent in 1971.[2] Respondents today are no longer accustomed to communicating by physical mail or answering surveys by telephone via landlines. Indeed, they do not always see a qualitative difference between Google Trends and the National Income and Product Accounts.

The National Academy of Sciences' Committee on National Statistics, first issued *Principles and Practices for a Federal Statistical Agency* in 1972 (the sixth edition is

[1] Robert Groves, as Director of the Census Bureau in 2011, said: "The current Census Bureau survey and census methods are unsustainable. Changes must occur in the acquisition of data and construction of statistical information for the Census Bureau to succeed" (http://directorsblog.blogs.census.gov/2011/09/08/the-future-of-producing-social-and-economic-statistical-information-part-i/; accessed on April 29, 2015).

[2] The first networked email message is generally attributed to Ray Tomlinson using ARPANET in 1971, though an earlier attempt to communicate (a login request) in 1969 is sometimes given the credit. [http://www.pbs.org/newshour/updates/internet-got-started-simple-hello/]

Citro 2017). The Committee has long recommended that the Federal Statistical System (FSS) evolve to meet the needs of its users. It is to the credit of the Census Bureau, in particular former Director Robert Groves, that it and its partner, the National Science Foundation (NSF), recognized the need of the FSS to adapt and evolve. In 2011, the Census Bureau and the NSF began a number of grants to academic institutions to collaborate with federal statistical agencies to help them do so, in an environment that married basic research activities to the applied needs of these governmental agencies.

II. A BRIEF HISTORY OF THE NSF-CENSUS RESEARCH NETWORK

The NSF-Census Bureau Research Network (NCRN) was established in 2011. NSF disseminated a call for proposals to create research nodes, each of which was to be staffed by teams of researchers conducting interdisciplinary research and educational activities on methodological questions of interest and significance to the broader research community and to the FSS, particularly the Census Bureau.[3] Funding for the NCRN came largely from the Census Bureau. The solicitation, peer review, selection, and grant oversight for the networks was done in partnership with the NSF, which also contributed funds, and performed the post-award administration and review.[4] After peer review of the proposals received in response to the call, the NSF made grant awards to eight nodes: Carnegie Mellon University, University of Colorado-Boulder joint with the University of Tennessee, Cornell University, Duke University joint with the National Institute of Statistical Science (NISS), University of Michigan-Ann Arbor, University of Missouri, University of Nebraska-Lincoln, and Northwestern University.[5] A second solicitation, to establish a Coordinating Office, led to an award to Cornell and Duke/NISS. Aggregate funding for the network was approximately $25.7 million. Initial awards were made in October 2011 for a 5-year period. Supplemental awards and no-cost extensions allowed parts of the network to be funded through September 2018.

A challenging task in describing the NCRN is to convey the distinction between basic

[3] See https://www.ncrn.info/nodes and the web sites of the nodes referenced therein for more details.

[4] In addition to standard progress reports to the NSF, the NCRN network as a whole was evaluated by a “reverse” site visit conducted by outside consultants to NSF in February 2015.

[5] Colorado-Tennessee and Northwestern are designated “small” nodes and the rest “medium” nodes, based on the level of funding. The Colorado-Tennessee node later added a principal researcher from Florida State University.

research with applications to the federal statistical system and research motivated by deep understanding of the goals and operations of that system. The network includes several investigators with decades of direct collaboration with the FSS. But it also includes many more scholars, from the agencies and from academia, who only recently have invested in understanding the uses of the statistical products as well as the methods used to produce them. This focus has produced innovative applications and new methodologies that are immediately applicable to current systems. It also advanced the NCRN goal of engaging new researchers – both experienced and at the start of their careers – in research relevant to the future of the FSS. By posing a wide range of federal statistical problems to potential grantees (see the Appendix) and not specifying the approaches, the Census Bureau and NSF hoped to encourage fresh and innovative approaches of broad applicability.

The activities to date have covered both fundamental and applied statistical research and have focused at least in part on the training of current and future generations of researchers in skills of relevance to surveys and alternative measurement of economic units, households, and persons. The results of "basic" research covered by this grant program are described in the more than 400 papers sponsored by the NCRN program and published online or in academic journals (see https://www.ncrn.info/documents/bibliographies for a complete list). Many of these research products have "applied" implications important to FSS agencies. This applied potential of NCRN research began to be realized with direct applications within FSS agencies, as described in Section V. Other important outcomes are training researchers in using and improving novel data products of the statistical agencies and in preparing them to move to employment in the federal statistical service as an employee after graduation. We discuss each of these in turn, as well as the way that the combination and interaction of these activities has created the basis for sustained innovation in federal statistics, both inside and outside the federal government.

The remainder of this paper will be in four parts. The next section discusses in brief some of the key research findings of the eight nodes, organized into six topics: (1) Improving census and survey data collection methods; (2) Using alternative sources of data; (3) Protecting privacy and confidentiality by improving disclosure avoidance; (4) Using spatial and spatio-temporal statistical modeling to improve estimates; (5) Assessing data cost and quality tradeoffs; and (6) Combining information from multiple sources. Later sections explore collaborations across nodes and with federal agencies, new software developed, and education activities and outcomes. The paper concludes with an evaluation of the ability of the FSS to apply the NCRN's research outcomes and suggests some next steps, as well as the implications of this research-network model for future federal government-academia collaborations.

III. SELECTED RESEARCH OF THE NCRN NODES

The call for proposals issued by the NSF focused on ten issues of interest to the FSS:[6]

1. Traditional concepts of family and households, as well as traditional concepts of economic units, are rapidly evolving.
2. Participation rates in sample surveys of households and economic units are declining.
3. The complexity of economic units is increasing, with multiple establishments, loose alliances, multiple lines of business, virtual spatial attributes, and highly dynamic structures.
4. Editing and imputation techniques commonly used in sample surveys currently have few evaluative frameworks that guide decisions on what approaches maximally reduce bias in final estimates.
5. Administrative records, when combined with survey data, may offer radically increased efficiencies in household and business surveys.
6. While public use datasets have greatly benefited quantitative research in the social sciences, the data are increasingly threatened by risk of inadvertent re-identification of sample members.
7. Small domain estimation using survey data offers the promise of greatly expanded useful estimates from sample surveys.
8. Cognitive and social psychological insights into respondent self-reports in social science research have reduced measurement errors.
9. The use of statistical models for large-scale descriptive statistics has advanced in important ways.
10. New approaches to disseminating census data to users are emerging, and new requirements for confidentiality protection will be required.

Not all of these ten issues produced proposals, much less relevant applied research. We focus on the network's contributions in six main areas, acknowledging that there is some overlap among them.

A. Improving Census and Survey Data Collection Methods

The core mission of the Census Bureau (and the most expensive), taking a census of the population every 10 years, is derived from the U.S. Constitution and began in 1790.[7] Collection

[6] For more details, see Appendix A, an excerpt from NSF solicitation 10-621.

[7] Article I, Section 2 includes this wording: "Representatives and direct Taxes shall be apportioned among

of data on businesses (specifically manufacturing) began in 1810; the economic censuses began in 1905.[8] It was not until the 1930s that the Census Bureau began to collect data based on a scientific sample of households or businesses; the Current Population Survey (CPS) was begun in order to collect information on unemployment during the Great Depression.[9] Given the importance of the mission, it is perhaps not surprising that a good deal of NCRN research focused on improving these data collection vehicles. This subsection focuses on how NCRN research has and will help the Census Bureau and other FSS agencies improve their data collection directly, but it also has important links to other sections (specifically alternate sources of data, and improving disclosure avoidance).

It is clear to both academic researchers and Census Bureau professionals that one important path to a less expensive decennial census in 2020 is through the use of more up-to-date technology. The traditional Census Bureau approach is being rethought, especially since there will be widespread use of online census forms. Such broad census design issues have been the focus of the Carnegie Mellon node in its interaction with Census Bureau researchers. NCRN research on the effects of different types of census errors on the resulting allocations of funds and representation that has taken place at the Northwestern node, described further in subsection III.E below, provides guidance on where to focus error-reducing resources. Stephen Fienberg (a principal investigator for the Carnegie Mellon University node) passed away before this paper could be drafted. Improving the Census was a touchstone of his career; his vision for the future of the Census is summarized in his 2013 Morris Hansen Lecture (Fienberg 2014).

The American Community Survey (ACS) is the second-most expensive item in the federal statistical budget.[10] Its *raison d'être* is the production of timely statistics for detailed

the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct." "Other Persons" meant slaves, emancipated by presidential proclamation in 1863 and subsequent constitutional amendments.

[8] https://www.census.gov/history/www/programs/economic/economic_census.html.

[9] https://www.census.gov/history/www/programs/demographic/current_population_survey.html. For more detail, see https://www.census.gov/prod/2002pubs/tp63rv.pdf, Chapter 2.

[10] In making these statements, we are not including the costs of collecting the administrative records that contribute to the statistical products of many agencies. Those costs are attributed to enforcing the

subpopulations, some of which, like geographic block-groups, are very small. In many ways, the ACS is the culmination of almost a century of refinement of large-scale sample surveys, beginning as noted with the CPS in the 1930s and the first sample "long-form" census in 1940.[11] The ACS is the prime example in the FSS of the principle that the most reliable information about how Americans live is gleaned from asking them directly.[12] There is a substantial body of scientific work that underpins reliance on scientific sampling, survey design, editing, quality control, and publication. But there is clearly a challenge in adapting these designs to the digital era – an era in which FSS access to administrative records can provide more accurate information that survey responses. That is where the integration of multimode surveys, including multiple online modes, and paradata analysis (analysis of data about the survey process) enters. The Carnegie Mellon, Duke, and Nebraska nodes have been very active in these areas. An additional challenge associated with the ACS is production of useful and interpretable small area estimates. The Missouri node has been active in the development of spatial and spatio-temporal model-based statistical methodology to improve estimates emanating from this survey (Bradley et al. 2015a, 2016a; Porter et al. 2014).

By studying survey data, paradata, and audio recordings, Nebraska node researchers have consistently found that the design of the questions plays a greater role in predicting survey data quality indicators (e.g., item non-response, response timing) and interviewer and respondent behaviors during a survey (e.g., exact question reading, provision of adequate answers) than characteristics of interviewers or respondents (Olson and Smyth 2015, 2017; Smyth and Olson 2015, 2016; Ganshert et al. 2016; Olson et al. 2016b; Sarwar et al. 2016; Timbrook et al. 2016a, 2016b, 2016c). This somewhat surprising finding speaks to the importance of question and questionnaire design. These researchers have also consistently found that respondent communicative and cognitive processes are affected by the interaction of respondents with interviewers within the overall context of the questions and their design features including calendar interviewing relevant for the Survey of Income and Program Participation (SIPP) (e.g.,

legislation that are the record systems' primary purposes.

[11] The ACS design and methodology is described in U.S. Census Bureau (2015).

[12] The principle of asking residents directly about their lives was encouraged in the first Congress by one of the Founding Fathers, Representative (later President) James Madison (see https://www.census.gov/schools/resources/historical-documents/james-madison.html).

Belli et al. 2013; Belli and Al Baghal 2016; Cochran and Smyth, 2014; Charoenruk 2015; Cochran et al. 2016; Olson et al. 2016a; Timbrook et al. 2016c; Kirchner and Olson 2017; Kirchner et al. 2017) and that interviewer behaviors and communicative processes are also affected by respondent behaviors and communicative processes (Timbrook et al. 2016c). Working with Gallup Panel data and paradata, Nebraska node researchers have reported that grids with higher complexity are more likely to induce data quality problems (Wang and McCutcheon 2016), and on the relationship between "Don't know" responses and survey satisficing (Wang et al. 2013) and the role of device types on survey quality (Wang et al. 2015). The findings from this research will be broadly applicable to federal statistical surveys.

Also related to questionnaire design and interviewer/respondent interactions, Nebraska researchers have suggested development of a prototype web-based computer-assisted telephone instrument for the Bureau of Labor Statistics (BLS) American Time Use Survey (ATUS) that features an intelligent agent that monitors interview progress and makes recommendations to the interviewer to help streamline data entry and improve the effectiveness and efficiency of interviewer-software interactions (Arunachalam et al. 2015; Atkin et al. 2015). The audit trails of the ATUS system were analyzed to identify respondent retrieval patterns and prompts were visualized to motivate respondents to answer relevant retrospective questions (Al Baghal et al. 2014; Atkin et al. 2014; Eck at al 2014).

One interesting corollary of an increased use of paradata in adaptive surveys is the necessity of organizing the storage, retrieval, and increased complexity in analytic tools needed for use of such data for analysis of large surveys (Olson and Parkhurst 2013), such as the multimodal ACS. Meanwhile, Nebraska node researchers have also applied data mining and machine learning techniques on paradata to predict respondent retrieval behaviors in calendar interviews (Belli et al. 2016) and respondent breakoffs (Eck et al. 2015a, 2015c; Wettlaufer et al. 2015; Eck and Soh 2017).

The Duke node has been working on understanding non-ignorable nonresponse; if a FSS agency can use stochastic editing to correct the data as they are collected and electronically transmitted, when should it stop data collection (a cost-benefit calculation)? Work by Paiva and Reiter (2017) creates a framework for that decision by modeling non-respondents and creating a simulation of the sensitivity of the results to further data collection. Clearly, non-respondents are different from respondents, but how different, and how does the agency determine whether they are different enough to affect the quantities under estimation? Other work at the Duke node could lead to significant improvements to editing and imputation throughout the FSS (Kim et al. 2015; Manrique-Vallier and Reiter 2016). Murray and Reiter (2016) present a nonparametric Bayesian joint model for multivariate continuous and categorical variables, with the intention of developing a flexible engine for multiple imputation of missing values. Their model fuses Dirichlet-process mixtures of multinomial distributions for categorical variables with Dirichlet-process mixtures of multivariate normal distributions for continuous variables, incorporating dependence between the continuous and categorical variables. White et al. (2016) adapt regression trees as engines for imputation of missing items in the Census of Manufactures, demonstrating how they improve on existing imputation routines for this central economic data product. Other relevant citations include Sadinle and Reiter (2017a, 2017b).

Once the data are collected, attention must be paid to making the data suitable for use. Editing the data for consistency to eliminate obvious errors (e.g., children older than their parents, pregnant males) is important, as is imputing for missing values (to make the public use data more user-friendly). The development of robust techniques integrating these methods (and applying them for the protection of confidentiality and the creation of synthetic data, described further below) has been a focus of the Duke node. For decades, FSS agencies have based their statistical editing practices on the principles elucidated by Fellegi and Holt (1976). Reiter and his colleagues have developed methods to improve on these time-honored methods by using Bayesian approaches to allow stochastic editing to create multiply imputed, plausible datasets. They propagate uncertainty about error localization – whereas traditional single imputation procedures lead researchers to underestimate uncertainty – and fully leverage information in the observed data to inform the edits and imputations. These developments include the use of these methods for both numerically valued economic data (Kim et al. 2015) and categorical-valued demographic data (Manrique-Vallier and Reiter 2016). Basically, these new techniques use statistical model-based approaches; the Census Bureau has begun a project to incorporate these methods into its 2017 Economic Censuses by using integrated edit, imputation, and confidentiality protection based on synthetic data models developed by Kim et al. (2015). They will permit publication of North American Product Classification System estimates and their margins of error without pre-specifying the table layout, as is currently done for the North American Industrial Classification System tabulations. This innovation illuminates how more accurate modern methods can substitute for less accurate but convenient historical ones.

B. Using Alternative Sources of Data

Censuses and surveys are not the only ways to collect information about the population and the economy. Independent sources can potentially provide useful data, such as from administrative records (AR) collected by governments for their own purposes (such as property assessments to levy real estate taxes or program applications to obtain benefits), and information provided by individuals in the course of their everyday activities (ranging from Twitter and Facebook posts to traffic-monitoring stations).

Making use of such information (particularly AR) in a statistical-agency environment typically requires record linkage, though there are cases where such information can be used without linkage (such as the Census Bureau's use of income tax records from the Internal Revenue Service for small businesses to avoid burdensome interviews). Record linkage is a critical component of the efforts to reduce census costs and, potentially, improve accuracy.[13]

Record linkage (or matching) occurs at virtually every stage of operational and experimental census designs:

- When a household address frame is the primary control system, record linkage occurs every time this frame is updated, primarily in the operation known as deduplication. The Census Bureau obtains a semiannual list of every address that the U.S. Postal Service delivers (or plans to deliver) mail and, after removal of commercial and governmental addresses, this list is used to update the Master Address File (MAF), which is used both to carry out a population and housing census and as a sampling frame for ongoing household surveys.
- When the operational frame is a master address list but the first decennial census contact is not from a traditional mail-in mail-back form, record linkage occurs when the

[13] Administrative records data have their limitations. As Groves and Harris-Kojetin (2017, p. 3-12) point out: "Administrative data can have many limitations including: (1) lack of quality control, (2) missing items or records (i.e., incompleteness), (3) differences in concepts between the program and what the statistical agency needs, (4) lack of timeliness (e.g., there may be long lags in receiving some or all of the data), and (5) processing costs (e.g., staff time and computer systems may be needed to clean and complete the data)."

responses are integrated as they are received, especially if they are received without a decennial census ID code.

Traditionally, the address on the mail-back form links directly to the MAF, linking the geography for the household to the accuracy of the MAF. When the first contact is via an online form (IP address) or cell phone (cellular location services), this information must be linked to the MAF. In the 2020 Census, Internet response can take one of two forms, called ID and non-ID. In the ID form, the respondent enters the encrypted MAF identifier supplied on the invitation to take the census. In the non-ID form, the respondent supplies a residential address directly. Processing the non-ID cases uses this alternative address information. Record linkage is expected to play a critical role in the non-ID processing. It will also likely play a critical role in the non-response follow-up stage via the use of information from multiple administrative record lists to complete the form in the absence of directly collected data (or supplementary to an incomplete report). Additionally, record linkage is one of an intruder's possible methods for attempting to break the confidentiality of released data, and thus one must assess the risk of confidentiality breaches from published tables and public-use microdata samples.

All of these (and other) record linkage applications can be quantitatively improved using new tools that simultaneously link more than two lists, while deduplicating each of the lists. This problem was addressed by the Carnegie Mellon, Duke, Michigan, and Cornell nodes in collaboration with researchers at the Census Bureau.[14] Their solutions provide conceptual generalizations of the familiar Fellegi-Sunter (1969) method for two lists (or deduplication of a single list) that are computationally feasible for application at the scale of the decennial census (Sadinle and Fienberg 2013, Steorts et al. 2013, Sadinle 2017). Further, the new methods acknowledge and propagate the uncertainty from the matching process into subsequent analyses. Improved record linkage can also improve the data needed to handle nonresponse to the census and to surveys, often by providing data for a particular address from AR, but also by providing the data for modeling non-respondents.

Particularly relevant for the Census Bureau is combining these issues into useful statistical models and methods. Fienberg (2015) presents a discussion of the value of addressing

[14] Researchers from all of these NCRN nodes and the Census Bureau formed multiple working groups to explore the implications of new record-linking technologies. The Summer Workgroup on Employer List Linking (SWELL), described further below, is one example.

(1) record linkage methods for three or more files, (2) combining duplicate detection and record linkage, (3) propagating duplicate detection and record linkage error into subsequent calculations, and (4) measuring both erroneous enumerations and omissions.

Record linkage is also important for business data. In collaboration with the University of Michigan's Sloan Foundation-funded Census-enhanced Health and Retirement Study (CenHRS), the Michigan node developed and tested methods for probabilistic linkage of the employers of HRS respondents to the Census Business Register (BR). This work addresses the complexity and benefits of linking household and business data to better understand employment of older Americans. The record linkage research confronts the difficulty of how individuals report their place of employment and how it is represented in administrative data. The approach taken highlights the importance of accounting for errors in matching records and of using probabilistic techniques to reflect these errors in subsequent analyses (Abowd et al. 2016).

The second alternative source of data for statistical agencies is "non-design data," also sometimes termed "organic data," "third-party data," "naturally-occurring data," or "data in the wild," such as from social media such as Twitter or transaction data that are digital traces of people's and businesses' daily activities (bank and credit card transactions, shopping, turning on lights, etc.). The key issue is not yet whether those data can replace data that FSS agencies use to report key social, economic, housing, and demographic indicators, but whether those data can provide useful indicators and checks on traditional time series, or produce measures at lower cost, greater frequency, more geographic detail, or in conjunction with survey data to reduce respondent burden.[15]

Account data. Data on consumers' transactions and balances can provide high-frequency and high-quality measures of spending, income, and assets that are difficult to measure accurately using surveys, which rely on infrequent self-reports from relatively small samples of individuals. In collaboration with the Sloan Foundation-funded database development project, the Michigan node pioneered the use of comprehensive account data from linked checking and credit card accounts to confront the difficulties of using such naturally-occurring account data to produce economically meaningful measurements and to study economic behavior and outcomes.

[15] Their use in official statistics could easily be jeopardized by changes in methodology by the independent provider, or even its discontinuation, as well as the proprietary nature of its collection and dissemination.

Gelman et al. (2014) show that account data drawn from a large sample of users of a financial services application can be broadly representative of the U.S. population. They use this newly developed data infrastructure to shed light on the excess sensitivity of spending to predictable income and use the same infrastructure to show how households accommodate short-run drops in liquidity (Gelman et al. 2015) and how their spending responds to a permanent change in gasoline prices (Gelman et al. 2016).

The use of transaction and balance data has great promise to improve spending and income measures published by the FSS. Spending reports are either based on very aggregate store-level data (the Census Bureau Advance Monthly Retail Trade Report) or surveys of consumers (the BLS Consumer Expenditure Survey). Both these surveys suffer from declining response rates and other data quality problems. Income reports when benchmarked to Internal Revenue Service tax data (such as the Bureau of Economic Analysis (BEA) National Income and Product Accounts and monthly Personal Income and Outlays) show survey underreporting. Tax data are inherently annual and available to the FSS only with a considerable lag and substantial limitations. Transaction data, on the other hand, are available daily, with high precision, for large samples of individuals, with great detail on location and type of spending, and with almost no lag.

Social media data. Official statisticians understand the framework in which a time series indicator like new unemployment insurance (UI) claims can be used to measure change. The population at risk is all statutory employees covered by state unemployment insurance systems. When the indicator goes down, fewer such employees filed new claims for UI. What does an increase in Tweets about "job loss" mean? The Michigan node developed a predictive model to assess this question. Job-loss Tweets do forecast the changes in official new claims for unemployment insurance (UI), particularly upward spikes, allowing one to capture turning points in economic activity that are often missed or captured only with a long lag using traditional approaches (Antenucci et al. 2013, 2014). The project developed a real-time predictor of UI claims and maintains a website giving weekly updates.[16]

An ongoing challenge to the use of social media data, in particular, for time series

[16] See econprediction.eecs.umich.edu. In 2015, the predicted trend of UI claims diverged from the actual, likely as a consequence of improving labor market conditions leading to reduced claiming.

measurement, is that while there is an enormous amount of this type of data in cross-section, no particular social media platform has existed long enough to capture an entire business cycle, let alone multiple such transitions. Thus, the development of measures from social media data, requires the systematic use of prior knowledge about the structure of the economy, such as how job flows change over the business cycle, akin to the use of seasonality adjustments. Without a benchmark reference, how does the predictive model detect a change in the weights it attaches to its inputs? The Michigan node is now addressing this issue with the development of an interactive model that allows those with domain expertise to provide benchmark datasets and economic concepts for measurement to a large archive of unstructured, web-based (social media and imaging) data in order to generate and archive new time series measures.

Researchers are also investigating natural-language processing of social media, transaction, and accounting data to help better understand economic measurement. The BLS, the Census Bureau, the BEA, and the Federal Reserve Board are watching this (and other related) research closely.

Other non-designed data. Another use of auxiliary data comes from combining area-level covariates measured over space and/or time with tabulated survey estimates within a hierarchical model-based framework. One salient example comes from the Missouri node's use of social media (functional time series) data from Google trends (Porter et al. 2014). The approach extends the traditional Fay-Herriot model to the spatial setting using functional and/or image covariates. A natural use for this methodology could be to incorporate remote sensing data as image covariates to augment information obtained from federal surveys or to assist with in-office address canvassing.

The Census Bureau, Michigan, and Cornell are collaborating to improve the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) database including investigations of its linkage to the SIPP. Additionally, the Michigan and Cornell nodes studied measurement issues for linked survey-administrative data, including HRS. Work at the Michigan node, has compared survey (SIPP) and administrative (LEHD) measures of the causes of job loss and has studied the implications for estimates of the response of earnings to job loss (Flaaen et al. 2016), developed and studied a measure of firm quality based on the ability of firms in the LEHD to attract and retain workers (Sorkin 2016), and developed explanations of the divergence of survey (HRS) and administrative (Social Security) measures of earnings (Hudomiet 2014). In work by the Cornell node and the Census Bureau, Green et al. (2017) investigated the coherence of survey and administrative reports of workplace location.

The Missouri node has proposed improvements to the statistics created from the LEHD database (Bradley et al. 2015a; 2017a) that make use of multivariate spatio-temporal statistical modeling. The Census Bureau and the Missouri and Cornell nodes are collaborating to enhance the precision of the disseminated estimates.

C. *Protecting Privacy and Confidentiality by Improving Disclosure Avoidance*

The NCRN researchers at Carnegie Mellon, Cornell, and Duke have long been involved in studying privacy and confidentiality.[17] Three different approaches to confidentiality protection span the ongoing work of the nodes in this area: data swapping (historically the Census Bureau method of choice to date for both the decennial census and the ACS),[18] multiple imputation (involving the preparation of synthetic datasets), and the more recently developed method of differential privacy that emanates from cryptography and computer science and offers the strongest possible privacy guarantees. However, differential privacy has not yet been proven to work for all kinds of data releases that the Census Bureau is accustomed to producing (Abowd and Schmutte 2016).[19] Recently, the *Journal of Privacy and Confidentiality* has devoted an entire issue (2015-2016, volume 7, issue 2) to differential privacy; see also Murray (2015-2016),

Both the Carnegie Mellon and Cornell nodes have contributed to the "the economics of privacy." Two key papers address their respective contributions. Acquisti et al. (2016) draw connections among diverse streams of theoretical and empirical research on the economics of privacy by focusing on the economic value and consequences of protecting and disclosing personal information, and on consumers' understanding and decisions regarding the trade-offs

[17] Privacy typically concerns what information a respondent is willing to share; confidentiality pertains to the ethical and statutory requirements to keep personal data from unauthorized disclosure to a third party.

[18] Methods akin to data swapping that have been used to perturb microdata to protect confidentiality include "blanking and imputing, data blurring, and a combination of micro agglomeration, substitution, subsampling, and calibration." (Groves and Harris-Kojetin 2017, p. 5-9.)

[19] In addition to the NCRN described in this article, the Census Bureau has also established cooperative agreements with Georgetown University and Purdue University to pursue additional research in disclosure avoidance and privacy. These two university groups participated with NCRN and Census Bureau researchers in an NSF-organized "Workshop on Practical Privacy" in the Fall of 2016 to focus on concrete problems of disclosure avoidance; see (Vilhuber et al. 2017).

associated with the privacy and the sharing of personal data (see also Acquisti et al. 2013 for some legal issues). Abowd and Schmutte (2017) consider the problem of determining the optimal accuracy of *public* statistics when increased accuracy requires a loss of privacy, and the role of who provides the statistics.

Acquisti et al. (2016) highlight how the economic analysis of privacy evolved over time, as advancements in information technology raised increasingly nuanced and complex issues. There are three themes they highlight: (1) Characterizing a single unifying economic theory of privacy is hard, because privacy issues of economic relevance arise in widely diverse contexts; (2) There are theoretical and empirical situations where the protection of privacy can both enhance and detract from individual and societal welfare; and (3) Consumers' ability to make informed decisions about their privacy is severely hindered because they are often in a position of imperfect or asymmetric information regarding when their data is collected, for what purposes, and with what consequences.[20]

But a much larger social issue also concerns researchers in the network. What are the appropriate tradeoffs between data confidentiality and data accuracy? As Abowd and Schmutte (2017) show, public statistics will be under-provided by private suppliers, and welfare losses from the under-provision can be substantial. But a key contribution of theirs is that the question cannot be answered from the technology of statistical disclosure limitation or privacy-preserving data mining. It requires understanding how the citizen consumers of the statistics an agency produces value data accuracy when they must pay with some loss of privacy. Statistical agencies are only beginning to come to grips with this question, but they are actually moving faster than large Internet firms such as Google and Facebook. All the players in this arena, public and private, understand the risks associated with direct privacy breaches far better than they understand how to measure a society's preferences for public data that can only be produced with some privacy loss. Changes to the current paradigm may require new legislation.[21]

Among the network's new contributions in this area is a focus on quantifying the disclosure risks associated with large-scale record linkage, such as that proposed for the 2020

[20] See also Acquisti et al. (2015).

[21] Title 13 U.S. Code, which covers the activities of the Census Bureau, prohibits any data release which might jeopardize confidentiality. All FSS agencies are covered by the Confidential Information Protection and Statistical Efficiency Act of 2002.

Census, and on producing accurate statistics that control that risk in a quantifiable way. Much of the NCRN research on disclosure avoidance addresses how to combine statistical disclosure limitation with correct analysis of the published data, including understanding the uncertainty introduced through probabilistic data linkage or model-based data imputation (Kim et al. 2017). Synthetic data with validation is one of the leading paradigms for researchers, especially when combined with data provenance curation (such as provided by the Cornell node's metadata work). The Duke node is collaborating with the Census Bureau on creating a synthetic-data version of the 2017 Census of Manufactures; there have been no public-use versions of the Economic Census microdata to date because of confidentiality concerns, particularly for outliers (e.g., very large firms).

An alternative to releasing data that have been aggregated, smoothed, perturbed, top- or bottom-coded, swapped, or otherwise massaged to protect confidentiality, is to produce either partial or fully synthetic data, where some or all of the data collected are replaced by draws from models designed to capture the distributional features of the collected data. The Cornell and Duke nodes have focused some attention on this approach, as has the Census Bureau, which currently produces some synthetic datasets for researchers. This approach has already caught the attention of the popular press (Callier 2015).

To the extent that the simulated data accurately reproduce the joint distribution of the variables of interest, the synthetic data can be used to make inferences. But given the difficulty of accurately reproducing all multivariate joint distributions present in the original (confidential) data, there needs to be a mechanism for validating such analyses and, if necessary, producing a new iteration of the synthetic dataset that corrects for anomalies. The Duke node has developed several techniques for providing users with feedback on the quality of their inferences from the synthetic data (Chen et al. 2017). Their measures satisfy differential privacy; an R software package is under development.

In the past, the Duke and Cornell nodes have focused on assisting the Census Bureau in the creation of a synthetic version of the Bureau's Longitudinal Business Database (Kinney et al. 2014). Business data present a much bigger challenge to maintain confidentiality than household data, given the presence of industry giants (Kinney et al. 2014, Miranda and Vilhuber 2016, Vilhuber et al. 2016). The Synthetic Longitudinal Business Database (SynLBD) is an experimental data product containing 21 million establishment records in all industry sectors for the years 1976-2000. The database contains the establishment's synthetic industrial classification code, its birth year, death year, annual payroll, annual employment, and whether it has more than one establishment, but it contains no geographic or firm-level information. As noted earlier, the site warns the user that unless the results are validated, there is no guarantee that results from the SynLBD reflect results from the underlying confidential data. Finally, the Missouri and Duke nodes have proposed disclosure avoidance methodology for spatially correlated data (e.g., see Quick et al. 2015, 2016).

One challenge addressed by research at the Cornell node was how to ensure that published scientific studies by internal FSS agency researchers or by outside researchers using the agency's confidential data could be audited and reproduced. As Lagoze et al. (2014) note, "Many of the data underlying social science research are embedded in complex provenance chains composed of inter-related private and publicly accessible data and metadata, multithreaded relationships among these data and metadata, and partially ordered version sequence, making it difficult to understand and trace the origins of data that are the basis of a particular study. This presents barriers to the essential scholarly tasks of testing research results for validity and reproducibility, creating a substantial risk of breach of the scientific integrity of the research process itself." Such basic scientific integrity measures require curation of the provenance of the data used in the studies. In turn, reproducibility of the use of confidential data ultimately improves its quality.

In Abowd et al. (2012), researchers at the Cornell node proposed enhancing one of the existing standards for curating metadata (the Data Documentation Initiative – DDI) in a way that respects all of the confidentiality constraints imposed on the curators (Lagoze et al. 2013a); the enhancement has been proposed to the body maintaining the standard. The fully curated metadata, including provenance, can be viewed and maintained within the confidential setting. A software system to demonstrate the enhancement, the Cornell Comprehensive Extensible Data Documentation and Access Repository (CED^{2}AR), is also provided (Lagoze et al. 2014). CED^{2}AR is in active use by the Census Bureau for the documentation of the SIPP "Synthetic Beta File" and the SynLBD, and is being considered for use elsewhere in the Bureau. In addition, it is being explored and developed in collaboration with the Inter-University Consortium for

Political and Social Research and the Roper Center for Public Opinion Research. Since it is based on a widely adopted standard, DDI, it should be considered an augmentation that will ease external access to and analysis of internal Census Bureau data, albeit in a protected environment (here, the Federal Statistical Research Data Centers, or FSRDCs).[22] Additional work aims to further expand the standard to embed provenance information, allowing researchers to tie diverse public-use and synthetic data products to common confidential source files (Lagoze et al. 2013b).

D. Using Spatial and Spatio-Temporal Statistical Modeling to Improve Estimates

The ACS design explicitly combines spatial and temporal information to produce annual and 5-year estimates for many subpopulations. These estimates are released with associated margins of error (MOEs, i.e., 90-percent confidence intervals). Working with current ACS data, researchers at the Missouri node and the Colorado-Tennessee node have each developed new spatial techniques for aggregating and disaggregating the basic ACS data geographically. In particular, research produced by the Colorado-Tennessee node focuses on the definitions of the spatial areas. Since the current practical approach to defining neighborhoods based on census tracts does not recognize population heterogeneity, the node has published open source data and software that allows users to create alternative ways to create spatial aggregates for which the statistics of interest can be estimated more accurately from the ACS data (Folch and Spielman 2014; Spielman and Folch 2015). These methods were motivated by interest in improving the usability of ACS estimates. The Colorado node documented geographic (and demographic) patterns in the quality of ACS estimates (Folch et al. 2016; Spielman et al. 2014) and found that most local government users of ACS data ignore estimate quality when using the data to make decisions (Griffin et al. 2014). These open source products have seen wide use and

[22] "FSRDCs are Census Bureau facilities, housed in partner institutions that meet all physical- and information-security requirements for access to restricted-use microdata of the agencies whose data are accessed at the FSRDCs. There are currently 23 FSRDCs, and they partner with more than 50 research organizations, including universities, nonprofit research institutions, and government agencies. Currently, four federal agencies (the Agency for Healthcare Research and Quality, the Census Bureau, the Bureau of Labor Statistics, and the National Center for Health Statistics) directly provide data through FSRDCs, and each agency has its own review and approval process. In addition to the agencies that directly provide their data, nine other agencies that sponsor surveys also participate in the FSRDC program by allowing surveys they cosponsor to be made available. In a further expansion of the role of FSRDCs, administrative data from other federal agencies are also being made more accessible to researchers through them." (Groves and Harris-Kojetin 2017, p. 5-12.)

commercialization (Spielman and Singleton 2015).[23]

In contrast, the Missouri node has developed a statistical framework for regionalization of multiscale spatial processes (Bradley et al. 2016a). The proposed method directly addresses the important modifiable areal unit and ecological fallacy problems associated with multiscale spatial data and introduces a criterion for assessing spatial aggregation error. This criterion, called CAGE (Criterion for spatial AGgregation Error), is then minimized to produce an *optimal* statistical regionalization. The impact of such methodology has significant implications for various FSS stakeholders. For example, various ACS data-users wishing to aggregate tabulations across geographies (using the methods discussed in Bradley et al. 2015b) can evaluate to what extent valid inferences can be made; an R software package is currently under development. This approach can be used to produce estimates on user-defined geographies and/or time. This could be used to obtain 3-year ACS period estimates (since discontinued) for regions that are not currently disseminated.

Results can be directly referenced to identifiable inputs in the statistical system and reproduced reliably from those inputs. Advances in the curation of the metadata help ensure that the agency's use of these methods can be audited and its published results can be reproduced. Reproducibility is not always possible for data science analysis based on commercial data such as Google Trends, but the Michigan node's research using Twitter feeds can be reproduced because they post their underlying data.

The Missouri node has been actively engaged in developing hierarchical statistical models that leverage different sources of dependence (e.g., multivariate, spatial, and spatio-temporal) to improve the precision of estimates from various data products. Broadly speaking, many of the proposed techniques can be viewed as natural generalizations of the methods currently used for small-area estimation by most statistical agencies—that is, they are generalizations of the Fay-Herriot (1979) model (e.g., Bradley et al. 2015a, 2015b, 2016a, 2016b; Porter et al. 2014, 2015a, 2015b, 2015c; Sengupta and Cressie 2013; Cressie and Zammit-Mangion 2016). The Missouri node has developed the hierarchical statistical-modeling approach in ways that will give federal statistical agencies a distinct advantage for their data products over commercial value-added resellers of the same data. This advantage stems directly

[23] See https://carto.com/data-observatory/

from the agency's access to and use of the complete set of geographic identifiers and original data values in doing the calculations, and then applying statistical disclosure limitation to the outputs (Quick et al. 2016). The methodologies developed at the Missouri node typically use the Census Bureau geography definitions, but they provide the flexibility to depart from this restriction. In other words, the proposed methods retain the ability to operate from customized geographies and/or temporal supports through the use of a change-of-support approach (Bradley et al. 2015b, 2016b).[24]

There are numerous examples of multiple surveys disseminating related demographic variables that are measured over space and/or time. The Missouri node's methodology combines the disseminated estimates from these surveys to produce estimates with higher precision. Additionally, in cases where estimates are disseminated with incomplete spatial and/or temporal coverage, the Missouri node's approach leverages various sources of dependence to produce estimates at every spatial location and every time point. The approach for combining the multiple surveys is developed as a fully Bayesian model. The proposed methodology is demonstrated by jointly analyzing period estimates from the Census Bureau's ACS and concomitant estimates obtained from the Bureau of Labor Statistics Local Area Unemployment Statistics program (Bradley et al. 2016a).

More generally, the Missouri node uses spatial, spatio-temporal, and/or multivariate dependence structures to generate point-in-time estimates of subpopulation quantities and provide an associated measure of uncertainty (wherein traditional small-area estimates are a special case). Flexible models have been introduced that allow estimation for both Gaussian and non-Gaussian settings (Sengupta and Cressie 2013; Bradley et al. 2015a, 2015b, 2016a, 2016b, 2017b, 2017c; Porter et al. 2014, 2015a, 2015b, 2015c). Extensions of the method can be used to incorporate other variables from the frame, or related frames. For example, Bradley et al. (2016a) introduces a multivariate mixed-effect spatio-temporal model that combines data from the Bureau of Labor Statistics' Local Area Unemployment Statistics with data from the ACS, to produce estimates that have significantly improved precision over using either survey

[24] To facilitate computation of space-time change-of-support, an R software package is being developed by the Missouri node in collaboration with Census Bureau research staff.

individually.

Visualization constitutes another important component in the analysis of spatial and spatio-temporal data. Using the ACS, Lucchesi and Wikle (2017) develop and present methods for simultaneously visualizing areal (spatial) data and its uncertainty using bivariate choropleth maps, map pixelation, glyph rotation, as well as animations. Spatial data can also be used to provide timely information about changing economic conditions. In work by the Michigan node that combines the themes of non-designed data and geo-spatial analysis, Wilson and Brown (2015) use Landsat imagery to show how the "Great Recession" affected southern Michigan by measuring changes in visible impervious surface area.

E. Assessing Data Cost and Quality Tradeoffs

Fundamental problems for the FSS (and for government statistical agencies around the world) include how to understand the value of the statistics they produce, how to compare value to cost in order to guide rational setting of statistical priorities, how to increase value for given cost, and how to better communicate the value of their data programs to those who set their budgets. The market does not provide a measure of value because government statistical data are public goods. So, to understand their value it is necessary to understand how the statistics are used, and what would occur if the statistics were available with different data quality characteristics. The Northwestern node extends and applies statistical decision theory, including cost-benefit analysis, to attack such basic questions.

The 2020 census of the U.S. is highly cost-constrained relative to previous censuses, and there is uncertainty about the quality of the census attainable for the allowed cost. Seeskin and Spencer (2015) considered alternative specifications of census quality and modeled the effects on (1) the funding allocation of perhaps $5 trillion over the decade of the 2020s, and (2) the distribution of seats in the U.S. House of Representatives in 2022. If the cost-quality relation can be specified, then their analysis permits estimation of the distortions in distributions of funds and seats that arise for given cost, in order to reveal the tradeoffs. For example, when the average standard deviation of a state's population is 2% of its actual population, the expected number of seats going to the wrong state is about 6.5, and the expected amounts of misallocated federal funds over the 10-year intercensal period is $40 billion. The expected absolute deviations in apportionments and in allocations both increased approximately linearly with the average relative standard deviation of state population numbers; for example, if the standard deviations of state populations doubles to 4% of actual, the expected number of malapportioned seats doubles to 13.[25]

In other work at the Northwestern node, Manski (2015) distinguishes transitory statistical uncertainty, permanent statistical uncertainty, and conceptual uncertainty. He illustrates how each arises as the BEA periodically revises Gross Domestic Product estimates, the Census Bureau generates household income statistics from surveys with nonresponse, and the BLS seasonally adjusts employment statistics. He anchors his discussion of communication of uncertainty in the contribution of Morgenstern (1963), who argued forcefully for agency publication of error estimates for official economic statistics.[26] In a related technical article, Manski (2016) elaborates on the theme of communicating uncertainty in official statistics, focusing on the permanent statistical uncertainty created by survey nonresponse. In current work, Manski is focusing on the crucial question of survey data collection design regarding how much data to collect and how much effort to expend to enhance the quality of the collected data when faced with a fixed budget. Dominitz and Manski (2017) use decision theory with a minimax regret principle for choosing between a high-cost high-accuracy survey and a low-cost low-accuracy one, where low accuracy is considered two ways – imprecise survey responses and unit non-response.

F. *Combining Information from Multiple Sources*

Distinguished from record linkage, which attempts to combine data sources in a way that matches information from multiple sources, better estimates can be made by combining information from multiple sources by modeling. One particular extant example is the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program.[27] The Missouri node has expanded this research field by developing a hierarchical Bayes approach using geography and/or time to enhance model estimation and prediction (Bradley et al. 2015b), in effect creating

[25] A cost-benefit analysis was conducted to compare the costs and benefits (in terms of fund allocation) for the 2016 quinquennial census of South Africa to an alternative of no census (Spencer et al. 2017).

[26] This has been done by the Census Bureau for decades for its monthly and quarterly economic indicators releases.

[27] The SAIPE program is described in detail at http://www.census.gov/did/www/saipe/. The program is an outgrowth of Fay-Herriot models originally built at the Census Bureau to produce per capita income estimates, discontinued in the late 1980s after the General Revenue Sharing Program (which used those estimates) was ended (Fay and Herriot 1979).

powerful spatio-temporal mixed effects models that include Fay-Herriot models as a special case. Given the available surveys, the conditional distributions of the latent processes of interest are used for statistical inference. To demonstrate the proposed methodology, researchers from the Missouri node have jointly analyzed period estimates from multiple surveys (Bradley et al. 2016a).

Other ways to improve socioeconomic estimates from the ACS involve models and data internal to the Census Bureau. For example, should modeling using external data sources be used to improve upon the direct survey estimates available from a household survey, and should survey-based (direct) estimates and model-based estimates, and indeed mixed (weighted) estimates, all be produced, or would confidentiality suggest limiting the types of data (and variables) that are modeled? The experience of the Census Bureau with its SAIPE and Small Area Health Insurance Estimates programs to address this question is relevant, as it attempts to expand the modeling to unemployment rates (noted above) and to the estimation of jurisdictions required to offer multi-lingual ballots under Section 203 of the Voting Rights Act.[28] Modeling can be used to generate new ACS estimates, other than those published for fixed geography and time periods (currently 1 year and 5 years), say a 4-year period estimate for a particular combination of census tracts representing a neighborhood (Bradley et al. 2015b).

IV. EDUCATIONAL OUTCOMES: STUDENTS, COURSEWARE, AND SOFTWARE

The NCRN nodes include faculty researchers, post-doctoral students, graduate students, and undergraduates. Each node worked actively with a designated coordinator at the Census Bureau, which also had a senior researcher to facilitate and coordinate its efforts.[29] More than 300 unique individuals authored or co-authored a paper listed on the NCRN Coordinating Office's consolidated bibliography. Each of the nodes presented in a "virtual" seminar, which was held roughly six times a year using multi-site distance-learning classrooms around the network and at the Census Bureau. Building on the popularity of statistics as an undergraduate major, nodes added official statistics components to both undergraduate and graduate courses, often as "special topics."

[28] See https://www.census.gov/rdo/pdf/3_VRA_Statistical_Methodology_Summary_V7.pdf.

[29] Daniel Weinberg from 2011 to 2014, then Nancy Bates assisted by Krista Park (briefly) and then Renee Ellis.

Cornell University (Abowd and Vilhuber) has led a distributed learning course, "Understanding Social and Economic Data," as a hybrid distance-learning/remote-learning course, with typical attendance involving a dozen sites, spread across the United States.[30] In 2013, more than 100 graduate students and faculty participated from 11 sites, including five NCRN nodes, the Census Bureau, and current and prospective FSRDC sites. The course was repeated in 2015 with about the same attendance. Course materials and video lectures are available online, which is serving as the basis for a Fall 2017 course offering. These efforts introduced a new generation of scholars to the FSS and its data. The Cornell node is considering turning the material into a textbook.

Nebraska created two new courses. The first course, "Interviewer-Respondent Interaction," explored different interviewing methods, methods to observe and analyze verbal behaviors during interviews, and methods to analyze these data.[31] The second course, "Survey Informatics," explored the role of technology throughout data collection, data management, and data analysis within survey research, as well as the increasing need for interdisciplinary teams within research to draw from the strengths of different disciplines (e.g., survey research and methodology, computer science and engineering, cognitive psychology, sociology, statistics, etc.) in order to properly answer interesting research questions and evaluate hypotheses Eck 2015).[32] An overview of Survey Informatics was published in Eck et al. (2015b). Further, a Nebraska student Charoenruk leveraged data produced by the Nebraska node to obtain an NSF Doctoral Dissertation Improvement Grant that studied how interviewer vocal characteristics in telephone surveys are related to data quality (Charoenruk 2015). Other NCRN-affiliated students completing their doctoral dissertations with NCRN support include Christine Chai of Duke University (Chai 2017) and Kristen Early of Carnegie Mellon University (Early 2017).

The University of Michigan offered a seminar for honors economics students, "Naturally-Occurring Data and the Macroeconomy" in 2016, wherein undergraduates did research using "big data" using techniques advanced by the Michigan node. This course will be offered in future years. The Michigan node supported independent research by doctoral students in economics. Aaron Flaaen used non-design data to create a new measure of the multi-national status of firms,

[30] http://www.vrdc.cornell.edu/info7470/

[31] http://digitalcommons.unl.edu/sociologyfacpub/490/

[32] http://digitalcommons.unl.edu/sociologyfacpub/489/

linked it to the Census Business Register, and made it available to Census Bureau researchers and researchers in the FSRDC network (Flaaen 2015); his analysis using these measures received the World Trade Organization Award for Young Economists. Isaac Sorkin developed and implemented a method for measuring employer quality based on the firm's relative ability to hire and retain employees. This work used eigenvalue techniques that allow analysis of flows across all connected establishments in the United States (Sorkin 2016). Peter Hudomiet developed explanations of the divergence of survey (HRS) and administrative (Social Security) measures of earnings (Hudomiet 2014).

One hope was that node-trained students would choose to work at a FSS agency upon graduation. Of course, successfully trained students also have other options, and it is difficult to assess empirically how many students gave the FSS consideration as an employment opportunity. As of this writing, we are aware of just two NCRN-trained graduates at the U.S. Census Bureau, from the Duke and Missouri nodes, though several students have accepted positions at other agencies and companies that interact closely with the FSS. Based on the authors' own experience in guiding students through the placement process, and based on interviews with colleagues and former students, a few tentative conclusions can be drawn. First, students do consider the agencies comprising the FSS as potential and attractive employers. However, in this era of "data science," the salary structure of the federal government is not competitive enough to attract such individuals. Furthermore, while graduate students are drawn from many countries, and NSF funding is available to international students, those same students cannot be hired by most federal agencies, due to legal restrictions that require an employee to be a U.S. citizen. Nonetheless, the exposure of such students to federal datasets and the challenges facing the federal statistical agencies still has some benefits. As these individuals either continue their education or go on to academic jobs, they take with them an appreciation for federal statistical problems and may continue to focus on federal statistics as research topics.

The nodes have developed short courses, workshops, and modules for use in college courses. These include:

- Short course on spatio-temporal statistics taught at the Census Bureau but open to staff at other FSS agencies (Missouri).

- Short course, “Introduction to Privacy” (Carnegie Mellon).[33]
- Short course on record linkage (data matching) (Carnegie Mellon).[34]
- Short course on missing data for the Odum Institute (Duke).
- Short course on synthetic data for the JPSM and the 2017 JSM (Duke).
- Topic modules on causes and statistical models for interviewer effects in survey data (Nebraska).
- Workshop on spatial demography and small-area estimation, “Measuring People in Place” at University of Colorado (Colorado-Tennessee).
- Workshops on using the SIPP and the synthetic SIPP (with matched earnings records from the Social Security Administration), conducted at Michigan, Duke, Census, and Population Association of America annual meetings, taught by Michigan and Census Bureau researchers (Michigan).[35]
- A 2-day workshop on Spatio-Temporal Design and Analysis for Official Statistics, organized and hosted by the Missouri node in May 2016. More than 40 researchers invited from both inside and outside the NCRN were involved in a series of break-out discussions. A summary of those discussions was distributed to workshop participants (Holan et al. 2016).

These educational activities have been particularly important in increasing usage of new, innovative Census data products that are in part the result of NCRN collaborations. For example, Census and the Cornell node worked closely to produce the synthetic SIPP, but the novelty of the data has limited its use to a small group of methodological researchers interested in synthetic data *per se*. A series of courses and the workshop described above introducing graduate students and junior scholars interested in studying the causes and consequences of poverty to the Survey of Income and Program Participation resulted in an increase in the number of users of the SIPP synthetic data and culminated in a panel on “Data Gold! Exploiting the Rich Research Potential of Lifetime Administrative Earnings Data Linked to the Census Bureau's Household SIPP Survey” at the 2016 joint American Social Science Associations-Labor and Employment

[33] http://www.stat.CMU/NCRN/PUBLIC/education.html#Priv
[34] http://www.stat.CMU/NCRN/PUBLIC/education.html#RLF13
[35] http://ebp-projects.isr.umich.edu/NCRN/training.html

Relations Association meeting.

The nodes have taken on the task of creating software for others to use in both improving and analyzing federal datasets.[36] Some illustrative software products include:

- The Colorado-Tennessee node developed and released open-source software for producing new statistical areas (out of existing census areas such as census blocks). This software reduces the variance in ACS estimates through intelligent aggregation.
- The Cornell node produced software to edit DDI-formatted metadata, called CED²AR. No existing DDI editor could show the additional features that Cornell had incorporated into the existing (DDI-C) standard, thus requiring the creation of the editor to be able to edit and display the additional data.
- The Duke node has developed several R software packages implementing missing data techniques, including the stochastic edit-imputation for continuous data of Kim et al. (2015), the model for mixed categorical and continuous data of Murray and Reiter (2016), the nonignorable imputation method of Paiva and Reiter (2017), and the model for categorical data with structural zeros of Manrique and Reiter (2014). It also developed software for generating synthetic values of the decennial census short form variables, using the methodology in Hu et al. (2017); the software ensures that structural zeros are respected (e.g., a daughter cannot be older than her biological father), and it captures within-household relationships.
- The Michigan node developed software in STATA and SAS, and a related STATA command, to improve the standardization of employer names and thereby improve record-linkage software for businesses (Wasi and Flaaen 2015). It also improved software to impute tax liability to household surveys that are not linked to administrative data in order to compute the Census Bureau's alternative poverty measure.
- The Missouri node is working on R software to implement customized geography and/or times. This software will automate the methodology of Bradley et al. (2015b). It is also collaborating with Esri on R software to quantify aggregation error from combining smaller geographies, allowing more efficient inferences (Bradley et al. 2017a).

[36] Links to the software listed below, and other software products, can be found at https://www.ncrn.info/software.

- The Missouri node has developed R code for visualizing the uncertainty in (spatial) areal data. This software appears in the online supplement to Lucchesi and Wikle (2017) and in the VizU R package available on Github.
- The Nebraska node has developed a program to automate scrubbing of computer-assisted survey audit trails to ensure confidentiality of all text fields, implemented at the Census Bureau, which enabled release of thousands of audit trails by replacing costly and time-consuming human intervention with automated processes.

V. THE IMPORTANCE OF COLLABORATION

As the NCRN matured, the opportunities and desirability of direct collaboration across the nodes and with the FSS agencies (particularly the Census Bureau) became more apparent. We focus first on inter-nodal collaborations, some of which resulted from movement of students between nodes (e.g., from post-doctoral fellow at one node to faculty member at another node). Examples of these collaborations include:

- Duke and Missouri on generating synthetic geographies.
- Duke and Carnegie Mellon on Fellegi-Sunter improvement.
- Duke and Cornell on creation of synthetic establishment data.
- Missouri and most other nodes at the "Workshop on Spatial and Spatio-Temporal Design and Analysis for Official Statistics," held in Columbia, MO on May 20-21, 2016.
- Michigan, Carnegie Mellon, Cornell, and Duke on evaluating methods for probabilistic linkage.
- Michigan and Cornell on implementing model-based probabilistic linkage for economic units, enhancing surveys with measures from administrative data, and evaluating quality of survey measures using administrative data.
- Michigan and Duke on training for using the Survey of Income and Program Participation.
- Nebraska and Carnegie Mellon on preliminary discussions regarding the development of an automated calendar for survey use.
- Colorado-Tennessee and Missouri on approaches to estimate margins of error on aggregated ACS estimates.

- Missouri and Cornell on spatio-temporal models for the LEHD program.

It is likely that many of these collaborations took place only because these universities were linked through the NCRN, especially through the biennial meetings convened by the NCRN Coordinating Office (mostly at the Census Bureau), since the topics chosen by the nodes did not overlap very much (a conscious decision by the NCRN program sponsors).

Some of the more important direct collaborations of node researchers with Census Bureau staff include the following:

- Development of a model to predict 2020 Census quality, as measured by the accuracy of the state population totals (Northwestern);
- Assessment of respondent comfort with geolocation of their home (Carnegie Mellon);
- Improvements in multiple file matching methods to aid the 2020 Census (Carnegie Mellon);
- Research to better understand residential mobility (Colorado-Tennessee);
- Imputations for missing business and demographic data (Duke);
- Development of methods for synthetic business data creation (Duke);
- Creation of a synthetic data version of the 2017 Economic Censuses (Duke);
- Improvements in confidentiality protection of demographic data (Cornell);
- Convening of the Summer Working Group for Employer List Linking (SWELL) (Michigan, Cornell);
- Participation in the Census Bureau's ACS Data Products Design working group (Colorado-Tennessee);
- Provision of advice on plans for 2020 Census operations, specifically on geographic targeting for the communications campaign, non-response follow-up, and coverage measurement (Colorado-Tennessee);
- Development of an imputation methodology for the Monthly Advance Retail Trade Survey (MARTS), development of model-based statistical methodology for in-office address canvasing, and implementation of space-time methodology using ACS data (Missouri);
- Provision of advice to Census Bureau staff on revising the ATUS user interface where SIPP-EHC navigation patterns are shown to be associated with data quality, which have

potential implications for interviewer training (Nebraska);

- Work with the Census Bureau's Center for Survey Measurement to assist with detecting measurement error through paradata (Nebraska).

One of the most active collaborations was SWELL. The purpose of this group, which included researchers from the Michigan, Carnegie Mellon, and Cornell nodes in active collaboration with Census Bureau staff, was to develop tools for linking person-level survey responses to employer information in administrative records files using probabilistic record linkage. Once the linkage was accomplished, there were four areas of potential payoff: (1) Production of a research-ready crosswalk between survey responses and administrative employer records including quality metrics to help users assess the probability that a particular link is correct; (2) Comparison of self-reported to administrative measures (e.g., location, earnings, firm size, industry, layoffs) enabling the enhancement of data quality by improving edits and imputations; (3) Creation of improved or new measures available to users without increasing respondent burden; and (4) Investigation of new research questions that could not be answered by either dataset alone (e.g., through creation of new variables and longitudinal outcomes or histories). The group has produced software (in SAS and STATA) for standardizing business names to allow improved linkages between survey reports of business names and administrative data from those employers (for the STATA version, see Wasi and Flaaen 2015). The research also helps to improve the Census Bureau's ability to design employer surveys that sample firms based on the composition of their employees, so that there can be better and more representative estimates of the characteristics of the employers of American workers. This successful collaboration was only possible because of the existence of an FSRDC at each location, allowing the sharing of data and research in real time. Despite the seasonality implied by its name, it is an ongoing collaboration.

There are still challenges for the transfer of the new technologies and approaches to practical implementation. Most likely to produce technology transfers is direct collaboration between Census Bureau staff and node researchers. Unfortunately, there have been few direct collaborations with Census Bureau researchers, which was not for want of trying by the nodes. The lack of direct collaboration was possibly due to the production pressures inherent in a FSS agency and the need for federal-government researchers to focus on practical, applied research.

The SWELL collaboration does demonstrate the value of collaboration between academics and FSS staff when there are common scientific goals, especially where these intersect with operational requirements of the FSS. On the geography front, researchers affiliated with the Colorado-Tennessee node are collaborating with the U.S. Geological Survey (Wood et al. 2015), Oak Ridge National Laboratory, and the U.S. Forest Service to improve their use of small-area data.

One possible amelioration of this lack of direct collaboration would be through co-location. Several individuals have attempted to take the results of their basic research and assist the Census Bureau in implementing their results by working on-site at the Census Bureau. One common approach has been for these individuals to become temporary federal employees, either through the Intergovernmental Personnel Act, Schedule A, or through summer student employment or fellowships (such as the dissertation fellowships) or the "Summer at Census" program. These include John Abowd and Lars Vilhuber from Cornell, Scott Holan from Missouri, Hang Kim and Jerry Reiter from Duke, and Kristen Olson from Nebraska. Ph.D students Zach Seeskin from Northwestern University, and Aaron Flaaen and Isaac Sorkin from the University of Michigan participated in the Census Bureau's dissertation mentorship program. Still others have become off-site collaborators, working on such projects as improving the American Time Use Survey time diaries collected by the Census Bureau for the BLS, improving the SIPP event history calendar for the Census Bureau, and revising the Census of Manufactures edit and imputation and data-dissemination strategies. Other topics that these "partially resident" researchers are working on include capture-recapture methodology (relevant for the estimation of census error), small-area estimation for the ACS and other surveys, improving editing and imputation for missing data, improving record-linkage practices allowing for uncertainty, implementing better storage paradigms for paradata, determining how to use paradata to identify problems, and improving the LEHD database. Other collaborations include matching the SIPP to the LEHD database (including development of a new-firm quality measure), improving the measurement of pension buyouts, SWELL, linking import-export data to the LBD and to non-Census Bureau data on multinationals to allow new types of research (available only to Census

Bureau and FSRDC researchers).

VI. LESSONS LEARNED

We note with pride that the network's outcomes have been numerous and valuable. The NCRN has been recognized with the *Statistical Partnerships among Academe, Industry and Government* 2017 award from the American Statistical Association "for addressing methodological questions of interest to the federal statistical system and training future generations to design, conduct, analyze and report official statistics." The network nodes have individually been productive, both in the basic and the applied research domains, with many publications, including in high-impact journals. Cross-node and government-university collaborations have occurred that probably would not have happened in the absence of a network, encouraged by the semi-annual open meetings at the Census Bureau.

Yet, improvements are desirable and possible. We believe that there are four valuable lessons that have been learned about government-academic research partnerships. First, better coordination between the agency and academic partners leads to more useful research outcomes. One suggestion is that "ways be found to facilitate not only the ability of academic scholars to spend time working within … government agencies, but also that key agency career researchers be encouraged and detailed to spend significant periods of time at the university-based research nodes where they can actively participate in the development of methodologies and basic science advances being pioneered there."[37] As noted above, the Census Bureau has already implemented part-time employment relationships, allowing the agency to bring the university-based researchers onto their agency teams directly. Moreover, better dissemination and communication across FSS agencies would facilitate greater utilization of other relevant research as well.

Should a similar government-academic partnership be pursued in the future, we encourage the government agencies to think about likely collaborations in advance.[38] The cross-fertilization that will result from academics working in close collaboration with government researchers will also further a second goal: enhanced technology transfer. It is not enough for

[37] *NCRN Reverse Site Visit Report*, direct communication from NSF to the NCRN Coordinating Office, February 14, 2015.

[38] It is also suggested that grant awards be made well ahead of the fall semester to facilitate the hiring of post-doctoral students by the grantees.

academics to invent new and useful methods if it is difficult for the relevant agencies to adopt those new methods. Adoption of several new techniques emanating from the NCRN nodes is well underway at the Census Bureau, due in large extent to those same researchers assisting the Census Bureau with the adoption.

Third, it is also important to think through the issues of academic access to confidential data in advance. While participating in the FSRDC program (then the Census Bureau RDC program) was not a requirement for a grant, all but one of the nodes without an RDC eventually joined that program and their research benefitted from access to restricted data.[39] The FSRDC program can be another useful though not required method to link researchers together. The FSRDCs could also provide a convenient way for Census Bureau staff to work in an academic setting for extended periods without losing touch with ongoing agency activities that might require access to confidential data. Furthermore, the FSRDC program can be used to link together collaborators from many locales, whether at the host academic institution or not.[40]

Fourth, the ability of FSS agencies to hire students trained as statisticians, whether through government-academic partnerships or otherwise, needs to be improved. Such students have skills most other potential hires do not, and hiring them can enhance the integration of research results into FSS practices. The main impediments are threefold: the hiring process is complex, the federal wage structure is often not competitive with the academic market, and many students are foreign nationals and therefore not eligible under current rules. Making this hiring process easier will require some ingenuity and supervisory attention, as federal human resources personnel sometimes do not recognize some of these students' classes as relevant. One mechanism to consider is a periodic virtual hiring seminar for data-oriented students, perhaps run jointly by FSS agencies under the auspices of the Office of Management and Budget Office of Statistical and Science Policy.

Finally, note the (inevitable) challenges of managing a network comprised of researchers from many disciplines spread across both academia and government. Breaking the disciplinary

[39] The lone exception, Carnegie Mellon University, has previously been a Census RDC node, but had let its participation lapse. Its researchers have the option of travelling a relatively short distance, to either Pennsylvania State University or Census Bureau headquarters, to access confidential data.

[40] This geographic collaboration with like-minded researchers may well be more fruitful than forced collaboration across nodes, since the nodes were deliberately chosen to have limited overlap in their research plans.

silos, to go to true cross-disciplinary research, is a challenge under any circumstances, and previous NSF-funded networks have certainly encountered the same challenges. Add to that the difficulty of bridging the gap between theory and practice, and various gaps in expectations between academic researchers and government practitioners, and it is clear that any such project can take a while to produce results. Moreover, the path from preliminary results to applied research is sometimes hard to execute, even if is a clear goal of the academic researcher. A key insight is to keep the network participants talking with one another; the NCRN's semi-annual meetings were more frequent than those of many other networks, and hence may have led to a faster convergence of ideas and language.

Overcoming the challenges to cross-disciplinary collaboration created a unique collaboration. NSF often recognizes the long-term aspect of creating effective collaborations when creating centers of excellence but these are not typically initiated in collaboration with a non-grant-making agency like the Census Bureau, and the budgetary intricacies of an NSF-agency collaboration are challenging. Nevertheless, any future attempt at creating a network similar in scale and breadth to the NCRN should consider addressing the budgetary issues for at least a 10-year horizon.

REFERENCES

Abowd, John M. and Ian M. Schmutte. 2016. "Economic Analysis and Statistical Disclosure Limitation." *Brookings Papers on Economic Activity* (Spring): 221-293.

Abowd, John M. and Ian M. Schmutte. 2017. "Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods." Labor Dynamics Institute Document 37; available at http://digitalcommons.ilr.cornell.edu/ldi/37/.

Abowd, John M., Lars Vilhuber, and William C. Block. 2012. "A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs." In J. Domingo-Ferrer and I. Tinnirello (eds.). *Privacy in Statistical Databases*. Springer Berlin Heidelberg 7556: 216-225.

Abowd, John M., Margaret C. Levenstein, Dhiren Patki, Ann Rodgers, Matthew D. Shapiro, and Nadi Wasi. 2016. "Developing Job Linkages for the Health and Retirement Study." Joint Statistical Meetings. Chicago, IL.

Acquisti, Alessandro, Leslie K. John, and George Loewenstein. 2013. "What is Privacy Worth?" *Journal of Legal Studies* 42(2): 249-274.

Acquisti, Alessandro, Laura Brandimarte, and George Loewenstein. 2015. "Privacy and Human Behavior in the Age of Information." *Science* 347(6221): 509-514.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman. 2016. "The Economics of Privacy." *Journal of Economic Literature* 54(2), 442–492.

Al Baghal, Tarek, Robert F. Belli, A. Lynn Phillips, and Nicholas Ruther. 2014. "What are You Doing Now? Activity Level Responses and Errors in the American Time Use Survey." *Journal of Survey Statistics and Methodology* 2(4): 519-537.

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro. 2013. "Ringtail: Feature Selection for Easier Nowcasting." *WebDB* 2013.

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro. 2014. "Using Social Media to Measure Labor Market Flows." NBER Working Paper No. 20010 (March).

Arunhachalam, Hariharan, Greg Atkin, Adam Eck, Doug Wettlaufer, Leen-Kiat Soh, and Robert F. Belli. 2015. "I Know What You Did Next: Predicting Respondent's Next Activity Using Machine Learning." *Proceedings of the 70th Annual Conference of the American Association for Public Opinion Research (AAPOR)*, Hollywood FL, May 14-17.

Atkin, Greg, Hariharan Arunachalam, Adam Eck, Leen-Kiat Soh, and Robert Belli. 2014. "Designing an Intelligent Time Diary Instrument: Visualization, Dynamic Feedback, and Error Prevention and Mitigation." *Proceedings of the 69th Annual Conference of the American Association for Public Opinion Research (AAPOR)*, Anaheim CA, May 15-18.

Atkin, Greg, Hariharan Arunachalam, Adam Eck, Doug Wettlaufer, Leen-Kiat Soh, and Robert F. Belli. 2015. "Using Machine Learning Techniques to Predict Respondent Type from A Priori Demographic Information." *Proceedings of the 70th Annual Conference of the American Association for Public Opinion Research (AAPOR)*, Hollywood FL, May 14-17.

Atrostic, B. K., Nancy Bates, Geraldine Burt, and Adrienne Silverstein. 2001. "Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights." *Journal of Official Statistics* 17(2): 209-226.

Belli, Robert F. and Tarek Al Baghal. 2016. "Parallel Associations and the Structure of Autobiographical Knowledge." *Journal of Applied Research in Memory and Cognition* 5: 150-157.

Belli, Robert F., Ipek Bilgen, and Tarek Al Baghal. 2013. "Memory, Communication, and Data Quality in Calendar Interviews." *Public Opinion Quarterly* 77: 194-219.

Belli, Robert F., L. Dee Miller, Tarak Al Baghal, and Leen-Kiat Soh. 2016. "Using Data Mining to Predict the Occurrence of Respondent Retrieval Strategies in Calendar Interviewing: The Quality of Retrospective Reports," *Journal of Official Statistics* 32(3): 579-600.

Bradley, Jonathan R., Scott H. Holan, and Christopher K. Wikle. 2015a. "Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics." *Annals of Applied Statistics* 9: 1761-1791.

Bradley, Jonathan R., Christopher K. Wikle, and Scott H. Holan. 2015b. "Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates," *STAT* 4: 255-270.

Bradley, Jonathan R., Scott H. Holan, and Christopher K. Wikle. 2016a. "Multivariate Spatio-Temporal Survey Fusion with Application to the American Community Survey and Local Area Unemployment Statistics." *STAT* 5: 224-233.

Bradley, Jonathan R., Christopher K. Wikle, and Scott H. Holan. 2016b. "Bayesian Spatial Change of Support for Count-Valued Survey Data with Application to the American Community Survey." *Journal of the American Statistical Association* 111: 472-487.

Bradley, Jonathan R., Christopher K. Wikle, and Scott H. Holan. 2017a. "Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error." *Journal of the Royal Statistical Society - Series B* 79: 815-832.

Bradley, Jonathan R., Scott H. Holan, and Christopher K. Wikle. 2017b. "Bayesian Hierarchical Models with Conjugate Full-Conditional Distributions for Dependent Data from the Natural Exponential Family." *Journal of the American Statistical Association* (under revision), arXiv:1701.0756

Bradley, Jonathan R., Scott H. Holan, and Christopher K. Wikle. 2017c. "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data." *Bayesian Analysis* forthcoming, arXiv:1512.07273.

Brick, J. Michael and Douglas Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." *Annals of the American Academy of Political and Social Science* 45(1): 36-59.

Callier, Vivienne. 2015. "How Fake Data Could Protect Real People's Privacy." *The Atlantic* (July 30); at https://www.theatlantic.com/technology/archive/2015/07/fake-data-privacy-census/399974/.

Chai, Christine Peijinn. 2017. "Statistical Issues in Quantifying Text Mining Performance." Doctoral Dissertation, Duke University.

Charoenruk, Nuttirudee. 2015. "Interviewer Voice Characteristics and Data Quality" Doctoral Dissertation, University of Nebraska-Lincoln.

Chen, Yan, Ashwin Machanavajjhala, Jerome P. Reiter, and Andres F. Barrientos. 2017. "Differentially Private Regression Diagnostics." *Proceedings of the IEEE International Conference on Data Mining 2016*, pp. 81-90.

Citro, Constance F. (ed.). 2017. *Principles and Practices for a Federal Statistical Agency: Sixth Edition*. Committee on National Statistics; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine. Washington, DC: The National Academies Press. https://doi.org/10.17226/24810.

Cochran, Beth, Kristen Olson, and Jolene D. Smyth. 2016. "Interviewer Influence on Interviewer-Respondent Interaction during Battery Questions." Paper presented at the annual meeting of the American Association for Public Opinion Research, Austin, TX, May 12-15.

Cochran, Beth and Jolene D. Smyth. 2014. "Hours or Minutes: Does One Unit Fit All?" Presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago, IL: November 21-22.

Cressie, Noel and A. Zammit-Mangion. 2016. "Multivariate Spatial Covariance Models: A Conditional Approach." *Biometrika* 103: 915-935.

Dominitz, Jeff and Charles F. Manski. 2017. "More Data or Better Data? A Statistical Decision Problem." *Review of Economic Studies* (forthcoming).

Early, Kristin. 2017. "Dynamic Question Ordering: Obtaining Useful Information While Reducing User Burden." Doctoral dissertation, Carnegie Mellon University.

Eck, Adam, L. Stuart, G. Atkin, Leen-Kiat Soh, Alan McCutcheon, and Robert Belli. 2014. “Making Sense of Paradata: Challenges Faced and Lessons Learned.” Presented at the annual meeting of the American Association for Public Opinion Research, Anaheim CA: May 15-18.

Eck, Adam, Leen-Kiat Soh, Allan L. McCutcheon, and Robert F. Belli. 2015a. “Predicting Breakoff Using Sequential Machine Learning Methods.” Presented at the annual meeting of the American Association for Public Opinion Research, Hollywood FL, May 14-17.

Eck, Adam, Leen-Kiat Soh, Kristen Olson, Allan L. McCutcheon, Jolene Smyth, and Robert F. Belli. 2015b. “Understanding the Human Condition through Survey Informatics.” *IEEE Computer* 48(11):110-114.

Eck, Adam, Leen-Kiat Soh, Allan L. McCutcheon, and Robert F. Belli. 2015c. “Predicting Survey Outcomes using Sequential Machine Learning Methods.” Presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago IL, November 20-21.

Eck, Adam. 2015. “Teaching Survey Informatics for the Future of Survey Research.” Presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago IL: November 20-21.

Eck, Adam, and Leen-Kiat Soh. 2017. “Sequential Prediction of Respondent Behaviors Leading to Error in Web-based Surveys” To be presented at the annual meeting of the American Association for Public Opinion Research, New Orleans LA: May 18-21.

Fay, Robert E., III and Roger A. Herriot. 1979. “Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.” *Journal of the American Statistical Association* 74(366): 269-277.

Fellegi, Ivan P. and D. Holt. 1976. “A Systematic Approach to Automated Edit and Imputation.” *Journal of the American Statistical Association* 71: 17-35.

Fellegi, Ivan P. and A.B. Sunter. 1969. “A Theory for Record Linkage.” *Journal of the American Statistical Association* 40: 1163-1210.

Fienberg, Stephen. 2014. “Envisioning the 2030 Census.” 23rd Annual Morris Hansen Lecture, Washington Statistical Society. January. Listed as the 2013 Hansen lecture at http://washstat.org/hansen/.

Fienberg, Stephen E. 2015. "Discussion [of Special Issue on Coverage Problems in Administrative Sources]." *Journal of Official Statistics* 31(3): 527-535.

Flaaen, Aaron B. 2015. “Multinational Firms in Context.” In “Essays on Multinational Production and the Propagation of Shocks.” Doctoral dissertation, University of Michigan.

Flaaen, Aaron, Matthew D. Shapiro, and Isaac Sorkin. 2016. “Reconsidering the Consequences of Worker Displacements: Firm versus Worker Perspective.” Unpublished working paper.

University of Michigan.

Folch, David C. and Seth E.Spielman. 2014. “Identifying Regions Based on Flexible User Defined Constraints.” *International Journal of Geographical Information Science* 28(1): 164–184. doi: 10.1080/13658816.2013.848986

Folch, David C., Daniel Arribas-Bel, Julia Koschinsky, and Seth E. Spielman. 2016. “Spatial Variation in the Quality of American Community Survey Estimates.” *Demography* 53(5): 1535–1554.

Ganshert, Amanda, Kristen Olson, and Jolene Smyth. 2016. The Effects of Respondent and Question Characteristics on Respondent Behaviors. Paper presented at the American Association for Public Opinion Research annual meeting, Austin TX, May.

Gelman, Michael, Shachar Kariv, Matthew D. Shapiro, Dan Silverman, and Steven Tadelis. 2014. “Harnessing Naturally Occurring Data to Measure the Response of Spending to Income.” *Science* 345 (11 July 2014): 212-215.

Gelman, Michael, Shachar Kariv, Matthew D. Shapiro, Dan Silverman, and Steven Tadelis. 2015. “How Individuals Smooth Spending: Evidence from the 2013 Government Shutdown Using Account Data.” NBER Working Paper 21025.

Gelman, Michael, Yuriy Gorodnichenko, Shachar Kariv, Dmitri Koustas, Matthew D. Shapiro, Dan Silverman, and Steven Tadelis. 2016. “The Response of Consumer Spending to Changes in Gasoline Prices.” NBER Working Paper 22969.

Green, Andrew S., Mark J. Kutzbach, and Lars Vilhuber. 2017. “Two Perspectives on Commuting: A Comparison of Home to Work Flows Across Job-Linked Survey and Administrative Files.” U.S. Census Bureau Center for Economic Studies Discussion Paper 17-34. https://ideas.repec.org/p/cen/wpaper/17-34.html

Griffin, Amy L., Seth E. Spielman, Jason Jurjevich, Meg Merrick, Nicholas N. Nagle, and David C. Folch. 2014. “Supporting Planners’ Work with Uncertain Demographic Data.” GIScience 2014 Uncertainty Workshop, September 2014, Vienna, Austria; http://cognitivegiscience.psu.edu/uncertainty2014/papers/griffin_demographic.pdf

Groves, Robert M. and Brian A. Harris-Kojetin (eds). 2017. *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy.* Washington DC: The National Academies Press; https://www.nap.edu/catalog/24652/innovations-in-federal-statistics-combining-data-sources-while-protecting-privacy.

Holan, Scott H. Noel Cressie, Christopher K. Wikle, Jonathan R. Bradley, and M. Simpson. 2016. *Summary* of “Workshop on Spatial and Spatio-Temporal Design and Analysis for Official Statistics.” NSF-Census Bureau Research Network report archived at Cornell University Library.

Hu, Jingchen, Jerome P. Reiter, and Quanli Wang. 2017. “Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data.” *Bayesian Analysis* forthcoming.

Hudomiet, Péter. 2014. “Dynamic Survey Measurement Error in Annual Earnings: The Role of Earnings Shocks, Cognition and Hidden Earnings.” In *Four Essays in Unemployment, Wage Dynamics and Subjective Expectations.* PhD Dissertation. University of Michigan.

Kim, Hang J., Jerome P. Reiter, and Alan F. Karr. 2017. "Simultaneous Edit-Imputation and Disclosure Limitation for Business Establishment Data." *Journal of Applied Statistics*.

Kim, Hang J., Lawrence H. Cox, Alan F. Karr, Jerome P. Reiter, and Quanli Wang. 2015. “Simultaneous Editing and Imputation for Continuous Data.” *Journal of the American Statistical Association* 110: 987-999.

Kinney, Satkartar K., Jerome P. Reiter, and Javier Miranda. 2014. “SynLBD 2.0: Improving the Synthetic Longitudinal Business Database.” *Statistical Journal of the International Association for Official Statistics* 30: 129-135.

Kirchner, Antje, Kristen Olson, and Jolene Smyth. 2017. “Do Interviewer Post-Survey Evaluations of Respondents Measure Who Respondents Are or What They Do? A Behavior Coding Study.” *Public Opinion Quarterly* forthcoming,

Kirchner, Antje and Kristen Olson. 2017. “Experience or Cooperation? Examining Changes of Interview Length over the Course of the Field Period.” *Journal of Survey Statistics and Methodology*. 5(1): 84-108.

Lagoze, Carl, William C. Block, Jeremy Williams, John Abowd, and Lars Vilhuber. 2013a. Data Management of Confidential Data.” *International Journal of Digital Curation* 8(1): 265-278.

Lagoze, Carl, Jeremy Willliams, and Lars Vilhuber. 2013b. “Encoding Provenance Metadata for Social Science Datasets.” In Emmanouel Garoufallou and Jane Greenberg (eds.) *Metadata and Semantics Research.* Communications in Computer and Information Science 390. Springer International Publishing, pp. 123-134.

Lagoze, Carl, Lars Vilhuber, Jeremy Williams, Benjamin Perry, and William C. Block. 2014. “CED^{2}AR: The Comprehensive Extensible Data Documentation and Access Repository.” *ACM/IEEE Joint Conference on Digital Libraries* (JCDL 2014), London, United Kingdom, 2014. DOI: 10.1109/JCDL.2014.6970178

Lucchesi, Lydia R. and Christopher K. Wikle. 2017. “Visualizing Uncertainty in Areal Data with Bivariate Choropleth Maps, Map Pixelation, and Glyph Rotation.” *Stat, doi: 10.1002/sta4.150.*

Manrique-Vallier, Daniel and Jerome P. Reiter. 2016. “Bayesian Simultaneous Edit and

Imputation for Multivariate Categorical Data." *Journal of the American Statistical Association*

Manski, Charles F. 2015. "Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern." *Journal of Economic Literature*, 53(3): 631-53.

Manski, Charles F. 2016. "Credible Interval Estimates for Official Statistics with Survey Nonresponse." *Journal of Econometrics,* 191: 293-301.

Miranda, Javier and Lars Vilhuber. 2016. "Using Partially Synthetic Microdata to Protect Sensitive Cells in Business Statistics." *Statistical Journal of the International Association for Official Statistics* 32(1): 69-80.

Morganstern, Oskar. 1963. *On the Accuracy of Economic Observations:* Second Edition. Princeton: Princeton University Press.

Murray, Jared S. 2015-2016. "Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering." *Journal of Privacy and Confidentiality* 7(1): 3-24.

Murray, Jared S. and Jerome P. Reiter. 2016. "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence." *Journal of the American Statistical Association* 111(516): 1466-1479.

Nancarrow, Clive, Julie Tinson, and Martin Evans. 2004. "Polls as Marketing Weapons: Implications for the Market Research Industry." *Journal of Marketing Management* 20(5-6): 639-655.

Olson, Kristen, Antje Kirchner, and Jolene D. Smyth. 2016a. "Do Interviewers with High Cooperation Rates Behave Differently? Interviewer Cooperation Rates and Interview Behaviors." *Survey Practice*. 9(2): 1-11.

Olson, Kristen, Jolene D. Smyth, and Amanda Ganshert. 2016b. "'During the LAST YEAR, Did You …': The Effect of Emphasis in CATI Survey Questions on Data Quality" Paper presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago, IL, November 18-19, 2016.

Olson, Kristen, and Bryan Parkhurst. 2013. "Collecting Paradata for Measurement Error Evaluation." Chapter 3 in Frauke Kreuter (ed.), *Improving Surveys with Paradata: Analytic Uses of Process Information*. New York: John Wiley and Sons. Pp. 43-72.

Olson, Kristen and Jolene D. Smyth. 2015. "The Effect of CATI Questionnaire Design Features on Response Timing." *Journal of Survey Statistics and Methodology* 3(3): 361-396.

Olson, Kristen and Jolene D. Smyth. 2017. "The Effect of Question Characteristics, Respondents and Interviewers on Question Reading Time and Question Reading Behaviors in CATI Surveys." Paper presented at the American Association for Public Opinion Research annual meeting, New Orleans, LA, May.

Paiva, Thais and Jerome P. Reiter. 2017. “Stop or Continue Data Collection: A Nonignorable Missing Data Approach to Continuous Data.” *Journal of Official Statistics* 33: 579-599.

Porter, Aaron T., Scott H. Holan, Christopher K. Wikle, and Noel Cressie. 2014. “Spatial Fay–Herriot Models for Small Area Estimation with Functional Covariates.” *Spatial Statistics* 10: 27-42.

Porter, Aaron T., Scott H. Holan, and Christopher K. Wikle. 2015a. “Bayesian Semiparametric Hierarchical Empirical Likelihood Spatial Models.” *Journal of Statistical Planning and Inference* 165: 78-90.

Porter, Aaron T., Scott H. Holan, and Christopher K. Wikle. 2015b. “Multivariate Spatial Hierarchical Bayesian Empirical Likelihood Methods for Small Area Estimation.” *STAT* 4: 108-116.

Porter, Aaron T., Christopher K. Wikle, and Scott H. Holan. 2015c. “Small Area Estimation via Multivariate Fay-Herriot Models with Latent Spatial Dependence.” *Australian and New Zealand Journal of Statistics* 57: 15-29.

Quick, Harrison, Scott H. Holan, Christopher K. Wikle, and Jerome P. Reiter, 2015. “Bayesian Marked Point Process Modeling for Generating Fully Synthetic Public Use Data with Point-Referenced Geography.” *Spatial Statistics* 14: 439-451.

Quick, Harrison, Scott H. Holan, and Christopher K. Wikle. 2016. “Generating Partially Synthetic Geocoded Public Use Data with Decreased Disclosure Risk Using Differential Smoothing.” *Journal of the Royal Statistical Society - Series A* (under revision), arXiv:1507.05529.

Sadinle, Mauricio. 2017. “Bayesian Estimation of Bipartite Matchings for Record Linkage.” *Journal of the American Statistical Association* 112: 600-612.

Sadinle, Mauricio and Fienberg, Stephen E. 2013. “A Generalized Fellegi-Sunter Framework for Multiple Record Linkage with Application to Homicide Record Systems.” *Journal of the American Statistical Association* 108(502): 385-397.

Sadinle, Mauricio and Jerome P. Reiter. 2017a. “Itemwise Conditionally Independent Nonresponse Modeling for Multivariate Categorical Data**.**” *Biometrika* 104(1): 207-220.

Sadinle, Mauricio and Jerome P. Reiter. 2017b. “Sequential Identification of Nonignorable Missing Data,” *Statistica Sinica* forthcoming.

Sarwar, Mazen, Kristen Olson, and Jolene D. Smyth. 2016. “Response Scales: Effects on Data Quality for interviewer Administered Surveys.” Paper presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago IL, November 18-19.

Seeskin, Zachary H. and Bruce D. Spencer. 2015. "Effects of Census Accuracy on Apportionment of Congress and Allocations of Federal Funds." Northwestern University, Institute for Policy Research Working Paper WP-15-05.

Sengupta, Aritra and Noel Cressie. 2013. "Hierarchical Statistical Modeling of Big Spatial Datasets Using the Exponential Family of Distributions." *Spatial Statistics* 4: 14 44.

Smyth, Jolene D. and Kristen Olson. 2015. "Recording What the Respondent Says: Does Question Format Matter?" Presented at the annual meeting of the American Association for Public Opinion Research, Hollywood, FL: May 14-17.

Smyth, Jolene D. and Kristen Olson. 2016. "How Do Mismatches Affect Interviewer/ Respondent Interactions in the Question/Answer Process?" Paper presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago IL, November 18-19.

Sorkin, Isaac. 2016. "Ranking Firms Using Revealed Preference." Unpublished paper, Stanford University.

Spencer, Bruce D., Julian May, Steven Kenyon, and Zachary Seeskin. 2017. "Cost-Benefit Analysis for a Quinquennial Census: The 2016 Population Census of South Africa.*" Journal of Official Statistics* 33.

Spielman, Seth E. and David C. Folch. 2015. "Reducing Uncertainty in the American Community Survey through Data-Driven Regionalization." *PLoS ONE* 10(2): e0115626. doi:10.1371/journal.pone.0115626

Spielman, Seth E., David Folch, and Nicholas Nagle. 2014. "Patterns and Causes of Uncertainty in the American Community Survey." *Applied Geography* 46: 147-157.

Spielman, Seth E. and Alex Singleton. 2015. "Studying Neighborhoods Using Uncertain Data from the American Community Survey: A Contextual Approach**.**" *The Annals of the Association of American Geographers* 102(5): 1003-1025.

Steorts, Rebecca C., Rob Hall, and Stephen E. Fienberg. 2016. "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association* 111(516): 1660-1672.

Timbrook, Jerry, Jolene D. Smyth, and Kristen Olson. 2016a. "Are Self-Description Scales Better than Agree/Disagree Scales in Mail and Telephone Surveys?" Poster presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago IL, November 18-19.

Timbrook, Jerry, Jolene D. Smyth, and Kristen Olson. 2016b. "Does Adding 'Your Best Estimate is Fine' Affect Data Quality?" Paper presented at the International Conference on Questionnaire Design, Development, Evaluation, and Testing, Miami FL, November 9-13.

Timbrook, Jerry, Jolene Smyth, and Kristen Olson. 2016c. “Why do Mobile Interviews Take Longer? A Behavior Coding Perspective.” Paper presented at the American Association for Public Opinion Research annual meeting, Austin TX. May.

Tourangeau, Roger and Thomas J. Plewes (eds.). 2013. *Nonresponse in Social Science Surveys: A Research Agenda.* Washington, DC: The National Academies Press.

U.S. Census Bureau. 2015. *Design and Methodology*. American Community Survey Technical Paper 67. http://www.census.gov/acs/www/Downloads/tp67.pdf.

Vilhuber, Lars, John M. Abowd, and Jerome P. Reiter. 2016. “Synthetic Establishment Microdata Around the World.” *Statistical Journal of the International Association of Official Statistics* 32(1); 65-68.

Vilhuber, Lars, Ian M. Schmutte, and John M. Abowd. 2017. “Proceedings from the 2016 NSF–Sloan Workshop on Practical Privacy." Cornell University Labor Dynamics Institute Document 33. http://digitalcommons.ilr.cornell.edu/ldi/33/

Wang, Mia and Allan McCutcheon. 2016. “Grids and Online Surveys: Do More Complex Grids Induce Survey Satisficing? Evidence from the Gallup Panel.” Presented at the annual meeting of the American Association for Public Opinion Research, Austin TX: May 12-15.

Wang, Mia, Leah Ruppanner, and Allan L. McCutcheon. 2013. “Do ‘Don’t Know’ Responses = Survey Satisficing? Evidence from the Gallup Panel Paradata.” Presented at the annual meeting of the American Association for Public Opinion Research, Boston, MA: May 16-19.

Wang, Mia, Allan L. McCutcheon, and Laura Allen. 2015. “Grids and Online Panels: A Comparison of Device Type from a Survey Quality Perspective”, Presented at the annual meeting of the American Association for Public Opinion Research, Hollywood FL: May 14-17.

Wasi, Nada and Aaron Flaaen. 2015. “Record Linkage Using Stata: Preprocessing, Linking, and Reviewing Utilities.” *Stata Journal* 15(3): 672-697.

Wettlaufer, Doug, Hariharan Arunachalam, Greg Atkin, Adam Eck, Leen-Kiat Soh, Robert F. Belli. 2015. “Determining Potential for Breakoff in Time Diary Survey Using Paradata.” *Proceedings of the 70th Annual Conference of the American Association for Public Opinion Research*, Hollywood FL, May 14-17.

White, T. Kirk, Jerome P. Reiter, and Amil Petrin. 2016. “Imputation in U.S. Manufacturing Data and Its Implications for Productivity Dispersion.” *Review of Economics and Statistics* forthcoming. Also published as National Bureau of Economic Research Working Paper #22569.

Wilson, Courtney R. and Daniel G. Brown. 2015. “Change in Visible Impervious Surface Area in Southeastern Michigan Before and After the ‘Great Recession’.” *Population and Environment*

36(3): 331-355.

Wood, Nathan J., Jeanne Jones, Seth Spielman, and Matthew C. Schmidtlein. 2015. “Community Clusters of Tsunami Vulnerability in the US Pacific Northwest.” *Proceedings of the National Academy of Sciences* 112(17): 5354-5359.

APPENDIX A. EXCERPT FROM NSF SOLICITATION 10-621 TO ESTABLISH THE NATIONAL SCIENCE FOUNDATION-CENSUS BUREAU RESEARCH NETWORK

Some questions currently of interest related to data collection, analysis, and dissemination processes include the following (these topics are not exhaustive):

Traditional concepts of family and households, as well as traditional concepts of economic units, are rapidly evolving.

- What methods can improve universe frame coverage of persons with intermittent ties with households, for entrepreneurial activities leading to new economic units in economic unit frames?
- What data auxiliary to households and covered persons might be used to estimate the propensity to be covered, as a targeting tool for alternative ways of assembling universe frames?
- Can theories be developed to guide research decisions for sampling unit definitions (derived from frames) and measurement units (e.g., enterprises vs. establishments, households vs. persons) to improve overall designs?
- How can estimates of immigration (both documented and undocumented) be improved?
- Is the concept of an "establishment" still relevant given changing business models and increasingly heterogeneous economic activity?

Participation rates in sample surveys of households and economic units are declining.

- What theories can inform the linkage between nonresponse rates and nonresponse errors?
- What data might be collected or linked to traditional survey data to improve the postsurvey adjustment for nonresponse to reduce nonresponse errors?
- What mechanisms underlie the finding that offering choices of alternative modes of data collection depress overall participation? What antidotes might be created to reduce that effect?
- How can administrative records on persons, households, and economic units be used in conjunction with traditional sample surveys to reduce the nonresponse error of traditional surveys?

The complexity of economic units is increasing, with multiple establishments, loose alliances, multiple lines of business, virtual spatial attributes, and highly dynamic structures.

- How can administrative records be used to improve the tailoring of measurement techniques to diverse types of economic units?
- How can changes in key attributes of economic units be tracked over time to improve the collection of data from the units?
- In longitudinal measurement, how can deaths, mergers, and acquisitions of economic units be forecasted to permit realtime measurement of those phenomena?
- How can multiple modes of data collection facilitate measurement of complex economic units?
- How can we more accurately classify heterogeneous economic activity within business enterprises, individual locations, or aggregates of locations?

Editing and imputation techniques commonly used in sample surveys currently have few evaluative frameworks that guide decisions on what approaches maximally reduce bias in final estimates.

- What logical or statistical approaches might offer guidance to the tradeoff decision of how much editing is optimal for diverse purposes?
- What editing algorithms might be developed to reduce the post-estimation review processes common in statistical estimation?
- What computer-assistance in editing might be developed to reduce the use of subject matter expertise in the review of data from longitudinal and other surveys?
- How can empirical diagnostic tools for evaluating auto-coding algorithms and large scale imputation approaches be improved?

Administrative records, when combined with survey data, may offer radically increased efficiencies in household and business surveys.

- What mathematical and statistical frameworks might be used to improve inference from probabilistically linked datasets?
- How can the social science community effectively monitor public attitudes toward administrative record usage?
- What conceptual frameworks might be developed to measure the error properties of linked survey and administrative record data?
- What imputation techniques can be created to deal with item missing data in linked files with variables common to multiple datasets?

While public use datasets have greatly benefited quantitative research in the social sciences, the data are increasing threatened by risk of inadvertent reidentification of sample members.

- What disclosure avoidance techniques can be developed to preserve pledges of confidentiality and maximize access to data?
- Can disclosure risk measurements be invented to guide practical decisions of data collectors regarding the release of data?
- How can synthetic data be produce that mimic the statistical properties of actual data but protect the identity of respondents?
- What effective analytic software approaches might be used to permit analysis of data without direct access to the data and protect pledges of confidentiality?

Small domain estimation using survey data offers the promise of greatly expanded useful estimates from sample surveys.

- How can model diagnostics be improved on small domain estimators?
- What small domain estimation approaches can exploit the longitudinal nature of surveys?
- What alternative approaches offer improved simultaneous estimation of small domains and higher level aggregates?
- What practical estimators of total error of small domain estimates might be developed for public dissemination?

Cognitive and social psychological insights into respondent self-reports in social science research have reduced measurement errors.

- What questionnaire development tools are superior for detecting different mechanisms of response error?
- What diagnostic tools in instrument development can be enhanced through computer assistance?
- How do we identify optimal measurement approaches for a single construct using individual modes of data collection?
- What diagnostics can be developed to isolate translation errors as a distinct component of measurement error in multilanguage measurement?

The use of statistical models for large-scale descriptive statistics has advanced in important ways.

- How can diagnostic tools be advanced to measure potential model-specification errors within a total error framework for the estimates?
- What diagnostic tools might be developed using model-based approaches to identify errors in tabular data?
- What models might be useful to estimate sampling error covariances and auto covariances in longitudinal estimates?
- What statistical models might be useful to forecast final estimates based on preliminary measurements of a sample?

New approaches to disseminating census data to users are emerging, and new requirements for confidentiality protection will be required.

- What metadata approaches will be most useful in documenting census data, and how can existing metadata systems be improved?
- How can census data dissemination, including both tabular and microdata, be improved?
- What are the most significant risks in disseminating census data to user communities, and how can those risks be diminished?
- What approaches can be developed that will allow the user community to safely and securely access census and other administrative data that have been merged across multiple agencies or sources?