NSF–Census Research Network (NCRN)

Spring 2017 Meeting

April 24, 2017

Washington DC

Location: U.S. Census Bureau HQ (Suitland Metro)

(the most up-to-date meeting agenda can always be found at https://www.ncrn.info/event/ncrn-meeting-spring-2017)

This version last updated 2017-04-19.

Anyone who does not have a badge will need to check in at the main gatehouse (across from the metro). They will have Renee Ellis' contact number. If you need anything please call Renee at 949-677-0196.

If you are not a US Citizen and do not have a Census Bureau badge, please register on Eventbrite using the "Foreign National" option at least 2 weeks before the conference - by close of business Wednesday, April 5th. If you have not registered by that time, entry will be denied.

You will need a security pass if you bring in a laptop- you can get those at the desk in the main entrance.

If you have a meeting and are unable to swipe the badge in the elevator to get to the floor you need, call Renee Ellis, she will assist you.

Monday, April 24, 2017

8:30-8:35 **Opening Remarks** - NCRN Coordinating Office - Lars Vilhuber

8:35-8:45 **Opening Remarks** - John Eltinge, Ass. Dir., Research and Methodology Directorate, Census Bureau

8:45-10:15 **Research Session I** [Census auditorium] **Linkage and Geography**

- Jared Murray, "Probabilistic Record Linkage after Indexing, Blocking and Filtering" (CMU) (25 minutes)
- David Folch, “Neighbors: The MAF provides new insights into the spatial organization of the American population” (University of Colorado at Boulder / University of Tennessee) (25 minutes)
- Matthew Simpson, “A Multiscale Spatial Approach to Change of Statistics” (University of Missouri) (25 minutes)
- Questions (15 minutes)

10:15-10:45 Break

10:45-11:45 **Research Session II:** [Census auditorium] **Looking forward**

- Carol Caldwell, “2017 Economic Census: Towards Synthetic Data Sets” (US Census Bureau) (25 minutes)
- Chris Clifton, "Practical Issues in Anonymity", (Purdue University) (25 minutes)
- Questions (10 minutes)

12:15-1:15 **PI-only Meeting (by invitation only)** [Conference room 3]
working lunch at Census Bureau (separate room, catered lunch)

1:30-2:30 **Research Session III** [Census auditorium] **Privacy and Confidentiality**

- Jerry Reiter, "Differentially Private Verification of regression model results" (Duke University) (25 minutes)
- Kobbi Nissim, "Formal Privacy Models and Title 13" (Georgetown University) (25 minutes)
- Questions (10 minutes)

4:00-5:00 **Meeting with Census Bureau (by invitation only)** [8H008 - Director's conference room]

6:30- NCRN Dinner (Lebanese Taverna) [registration required]

Abstracts

Research Session I: Linkage and Geography

Jared Murray, "Probabilistic Record Linkage and De-duplication after Indexing, Blocking, and Filtering"

Probabilistic record linkage, the task of merging two or more databases in the absence of a unique identifier, is a perennial and challenging problem. The number of possible links grows rapidly in the size of the databases under consideration, and in most applications it is necessary to first reduce the number of record pairs that will be compared. Spurred by practical considerations, a range of methods have been developed for this task. These methods go under a variety of names, including indexing and blocking, and have seen significant development. However, methods for inferring linkage structure that account for indexing, blocking, and additional filtering steps have not seen commensurate development. I describe the implications of indexing, blocking and filtering within the popular Fellegi-Sunter framework, and propose a new model to account for particular forms of indexing and filtering.

David Folch, "Neighbors: The MAF provides new insights into the spatial organization of the American population"

For decades the census tract has been the unit of choice for social scientists studying neighborhood context in the U.S. While few researchers would argue that census tracts match residents' conceptions of their neighborhood, the difficulty of developing alternatives, particularly for analyses covering multiple cities or regions, has kept the tract in broad use. Relatively recently, the Master Address File (MAF) was added to the Federal Statistical Research Data Center (FSRDC). The MAF provides a latitude/longitude point on nearly all respondents to the decennial census and the American Community Survey (ACS). Using the MAF, we have constructed egocentric neighborhoods where each respondent is the center of their own personalized neighborhood. This framework allows us to compare an individual's egocentric neighborhood to their tract to see how representative the tract is of their experience; and to compare it to the egocentric neighborhoods of people from different racial and economic groups. This work opens the door to understanding where census tracts work well, where they do not and strategies for considering alternative spatial aggregations of the population.

Matthew Simpson, "A Multiscale Spatial Approach to Change of Statistics"

Statistical agencies often publish multiple data products emanating from the same survey. First, they produce tabulated aggregate estimates of various features of the distributions of a several socio-demographic quantities of interest, including moments and quantiles. Often times, these area-level estimates are tabulated at small geographies. Second, statistical agencies frequently produce weighted public-use microdata samples (PUMS) that provide detailed information of the entire distribution of the same socio-demographic variables of interest. However, as these data are released at the unit level, the public-use micro areas usually constitute larger geographies (in terms of population size), in order to protect against the identification of households or individuals included in the sample. These two different data products represent a tradeoff in official statistics: publicly available data products can either provide detailed spatial information at the area level or detailed distributional information at the unit level within a larger area, but not both, as geocoded microdata are not released. We propose a model-based method to combine these two disparate data products to produce estimates of detailed features of a variable of interest at a high degree of spatial resolution. Our working example uses tabular estimates and PUMS from the American Community Survey to estimate U.S. Census tract-level income distributions and statistics of these distributions, not currently released, such as the Gini coefficient.

Research Session II: Looking forward

Carol Caldwell, "2017 Economic Census: Towards Synthetic Data Sets"

The U.S. Census Bureau conducts an Economic Census every five years, producing key measures of American business and the economy. Basic measures for all covered industrial sectors include number of establishments; sales, shipments, receipts, or revenue; annual and first quarter payroll; and mid-March employment. These general statistics items are highly correlated, and administrative data are available for modeling – and in many cases – validation. In addition, the Economic Census collects detailed information on the revenue obtained from products from all sampled units. The collection is quite challenging for a variety of reasons; legitimate missing values occur frequently and nonresponse is quite high. In contrast to the general statistics items collected from all establishments, auxiliary data on products are not readily available. Moreover, other predictors such as total receipts are often weakly related. Beginning in 2017, the Economic Census will begin collecting data electronically. In addition, the Economic Census will use the North American Product Classification System (NAPCS) to produce economy-wide product tabulations from cross-sector collections. By design, we expect the collected data to differ from the historic data collection. We do not expect that these changes will have substantive effects on the release of the tabulations of general statistics items. However, we suspect that many previously-released product estimates will be suppressed for either reliability or disclosure avoidance reasons.

This presentation discusses a planned collaboration between several departments at the U.S. Census Bureau and the University of Cincinnati to develop synthetic data sets from a subset of Economic Census industries (all sectors represented) comprising selected general statistics items and selected project value estimates using methodology developed by Kim, Reiter, and Karr (2016). The proposed approach satisfies all of the formal editing rules and includes formal privacy restrictions. The methodology was vetted on historic Economic Census data from the manufacturing sector for slightly different set of items, and has achieved satisfactory results. The primary purpose of the project is as "proof of concept," providing an opportunity to expand the method's modeling capabilities to include other parametric models and to test the model's robustness. The synthetic datasets will satisfy predetermined utility and privacy conditions and can be released to the public as a data product.

Chris Clifton, “Practical Issues in Anonymity”

Anonymization has been an ongoing topic of academic research, with a wide range of results in both techniques for anonymization and the privacy limits of those techniques. In our research, we have discovered a number of practical issues that fall outside of typical research, particularly with the utility of anonymized data (both public use microdata sets, and with privatized analysis results (e.g., differential privacy). This talk will discuss some of these issues, including scaling problems and utility. Specifically, we will introduce data partitioning for anonymization, issues with domain hierarchies, and choice of modeling techniques as areas that research largely considers "implementation details", but can have substantial impact on real-world quality of outcomes.

Research Session III: Privacy and Confidentiality

Jerry Reiter, “Differentially Private Verification of regression model results”

With growing risks of unintended disclosures, some agencies are turning to disclosure control methods that do not provide access to the original data, such as fully synthetic data or remote access (with output perturbation). Unfortunately with these methods it is difficult for users to know whether or not to trust the results of analyses. A user's model may fit the synthetic data, but poorly fit the confidential data; the (noisy) output might come from a model that fails to fit the data. I describe ways that agencies can provide feedback on the quality of analyses while controlling the information leaked by that feedback. Specifically I describe several quality measures that satisfy differential privacy.