NSF–Census Research Network (NCRN)

Spring 2016 Meeting

May 9-10, 2016

Washington DC

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

(the most up-to-date meeting agenda can always be found at http://www.ncrn.info/event/ncrn-spring-2016-meeting)

This version last updated 2016-04-22.

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

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, she will get you where you need to go.

Monday, May 9, 2016

10:00-10:15 **Opening Remarks** - NCRN Coordinating Office - Lars Vilhuber

10:15-10:30 **Opening Remarks** - John H. Thompson, Director, Census Bureau

10:30-12:00 **Research Session I** [Conference rooms 1-2] (Organizer: Lars Vilhuber)

- Duke: "*Itemwise missing at random modeling for incomplete multivariate data*" (**Mauricio Sadinle** and Jerry Reiter) (30 minutes)
- CMU: *"Assessing Respondent Attitudes Towards Geolocation in Online Surveys"* **(Laura Brandimarte**) (30 minutes)
- Nebraska: "*The ATUS and SIPP-EHC: Recent Developments*" (**Robert F. Belli**) (30 minutes)

12:00-1:00 **PI-only Meeting**, working lunch at Census Bureau (separate room, catered lunch)

Parallel:

1:00-4:00 Independent meetings with Census Bureau staff

1:25-4:00 INFO7470 final (live) session on Synthetic Data [T10]

4:00-4:30 **Meeting with Census Bureau** Director, Deputy Director, Associate Director R&M, staff

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

Tuesday, May 10, 2016

9:00-10:00 **Research Session II** [Conference rooms 1-2] (Organizer: Lars Vilhuber)

- Northwestern: *"A 2016 View of 2020 Census Quality, Costs, Benefits"* (**Bruce Spencer**)
- Nebraska: "*Data quality in time diary surveys*" (**Ana Lucía Córdova Cazar**)

10:00 Break

10:15-11:15 **Research Session III** [Conference rooms 1-2] (Organizer: Lars Vilhuber)

- Cornell: *"The Advantages and Disadvantages of Statistical Disclosure Limitation for Program Evaluation*" (**Ian Schmutte**)
- Michigan: "*Developing job linkages for the Health and Retirement Study"* (**Maggie Levenstein**)

11:15 Break

11:30-12:00 **Research Session IV** [Conference rooms 1-2] (Organizer: Lars Vilhuber)

- Cornell: "*Crowdsourcing Codebook Development and Enhancements in CED²AR – Progress on metadata"* (**Lars Vilhuber**, Bill Block)

End of meetings.

Abstracts: Monday

Mauricio Sadinle and Jerry Reiter: "Itemwise missing at random modeling for incomplete multivariate data"

Modeling multivariate data that are subject to missingness requires making assumptions about how the missing data arise. We introduce the concept of the missing data being itemwise missing at random (IMAR) when each random variable is conditionally independent of its missingness indicator given the remaining variables and their missingness indicators. We show that this assumption leads to a non-parametric saturated class of models and illustrate how to use it with a number of examples. We also show how to perform sensitivity analysis and explore how the IMAR assumption can be relaxed using marginal information from auxiliary sources.

Laura Brandimarte: "Assessing Respondent Attitudes Towards Geolocation in Online Surveys"

Geolocation refers to the automatic identification of the physical locations of Internet users. In an online survey experiment, we studied respondent reactions towards different types of geolocation. After coordinating with US Census Bureau researchers, we designed and administered a replica of a census form to a sample of respondents. We also created slightly different forms by manipulating the type of geolocation implemented. Using the IP address of each respondent, we approximated the geographical coordinates of the respondent and displayed this location on a map on the survey. Across different experimental conditions, we manipulated the map interface between the three interfaces on the Google Maps API: default road map, Satellite View, and Street View. We also provided either a specific, pinpointed location, or a set of two circles of 1- and 2-miles radius. Snapshots of responses were captured at every instant information was added, altered, or deleted by respondents when completing the survey. We measured willingness to provide information on the typical Census form, as well as privacy concerns associated with geolocation technologies and attitudes towards the use of online geographical maps to identify one's exact current location.

Robert F. Belli: "The ATUS and SIPP-EHC: Recent Developments"

One of the main objectives of the NCRN award to the University of Nebraska node is to investigate data quality associated with timeline interviewing as conducted with the American Time Use Survey (ATUS) time diary and the Survey of Income and Program Participation event history calendar (SIPP-EHC). Specifically, our efforts are focused on the relationships between interviewing dynamics as extracted from analyses of paradata with measures of data quality. With the ATUS, our recent efforts have revealed that respondents differ in how they handle difficulty with remembering activities, with some overcoming these difficulties and others succumbing to them. With the SIPP-EHC, we are still in the initial stages of extracting variables from the paradata that are associated with interviewing dynamics. Our work has also involved the development of a CATI time diary in which we are able to analyze audio streams to capture interviewing dynamics. I will conclude this talk by discussing challenges that have yet to be overcome with our work, and our vision of moving forward with the eventual development of self-administered timeline instruments that will be respondent-friendly due to the assistance of intelligent-agent driven virtual interviewers.

Abstracts: Tuesday

Bruce D. Spencer: "A 2016 View of 2020 Census Quality, Costs, Benefits"

Census costs affect data quality and data quality affects census benefits. Although measuring census data quality is difficult enough ex post, census planning requires it to be done well in advance. The topic of this talk is the prediction of the cost-quality curve, its uncertainty, and its relation to benefits from census data.

Ana Lucía Córdova Cazar: "Data quality in time diary surveys"

Over the past decades, time use researchers have been increasingly interested in analyzing wellbeing in tandem with the use of time (Juster and Stafford, 1985; Krueger et al, 2009). Many methodological issues have arose in this endeavor, including the concern about the quality of the time use data. Survey researchers have increasingly turned to the analysis of paradata to better understand and model data quality. In particular, it has been argued that paradata may serve as proxy of the respondents' cognitive response process, and can be used as an additional tool to assess the impact of data generation on data quality. In this presentation, data quality in the American Time Use Survey (ATUS) will be assessed through the use of paradata and survey responses. Specifically, I will talk about a data quality index I have created, which includes measures of different types of ATUS errors (e.g. low number of reported activities, failures to report an activity), and paradata variables (e.g. response latencies, incompletes). The overall objective of this study is to contribute to data quality assessment in the collection of timeline data from national surveys by providing insights on those interviewing dynamics that most impact data quality. These insights will help to improve future instruments and training of interviewers, as well as to reduce costs.

John M. Abowd and Ian M. Schmutte: "The Advantages And Disadvantages Of Statistical Disclosure Limitation For Program Evaluation"

This paper formalizes the manner in which statistical disclosure limitation (SDL) hinders empirical research in economics. We also highlight a hitherto unappreciated advantage of SDL, formal privacy models, and synthetic data systems: they can serve as a defense against model overfitting and false-discovery bias. More specifically, a synthetic data validation system can – and we argue should – be
used in conjunction with systems in which researchers register their research design ahead of analysis. The key insight is that privacy-protected data can be used for model development while minimizing risk of model overfitting. To demonstrate these points, we develop a model in which the statistical agency collects data from a population, but publishes a version in which the data that have been intentionally distorted by some SDL process. We say the SDL process is ignorable if inferences based on the published data are indistinguishable from inferences based on the unprotected data. SDL is rarely ignorable. If the researcher has knowledge of the SDL model, she can conduct an SDL-aware analysis that explicitly corrects for the effects of SDL. If, as is often the case, if the SDL model is unknown, we describe circumstances under which SDL can still be learned.

John Abowd, Margaret Levenstein, Kristin McCue, Ann Rodgers, Matthew Shapiro, Nada Wasi: Developing job linkages for the Health and Retirement Study

This paper documents work using probabilistic record linkage to create a crosswalk between jobs reported in the Health and Retirement Study (HRS) and the list of workplaces on Census Bureau's Business Register. Matching job records provides an opportunity to join variables that occur uniquely in separate datasets, to validate responses, and to develop missing data imputation models. Identifying the respondent's workplace ("establishment") is valuable for HRS because it allows researchers to incorporate the effects of particular social, economic, and geospatial work environments in studies of respondent health and retirement behavior. The linkage makes use of name and address standardizing techniques tailored to business data that were recently developed in a collaboration between researchers at Census, Cornell, and the University of Michigan. The matching protocol makes no use of the identity of the HRS respondent and strictly protects the confidentiality of information about the respondent's employer. The paper first describes the clerical review process used to create a set of human-reviewed candidate pairs, and use of that set to train matching models. It then describes and compares several linking strategies that make use of employer name, address, and phone number. Finally it discusses alternative ways of incorporating information on match uncertainty into estimates based on the linked data, and illustrates their use with a preliminary sample of matched HRS jobs.

Benjamin Perry, Venkata Kambhampaty, Kyle Brumsted, Lars Vilhuber, & William C. Block: Crowdsourcing Codebook Development and Enhancements in CED²AR

Recent years have shown the power of user-sourced information evidenced by the success of Wikipedia and its many emulators. This sort of unstructured discussion is currently not feasible as a part of the otherwise successful metadata repositories. Creating and augmenting metadata is a labor-intensive endeavor. Harnessing collective knowledge from actual data users can supplement officially generated metadata. As part of our Comprehensive Extensible Data Documentation and Access Repository (CED²AR) infrastructure, we demonstrate a prototype of crowdsourced DDI on actual codebooks. While the system itself is more general, the demonstrated implementation relies on a set of linked deployments of the basic software on web servers. The backend transparently handles changes, and frontend has the ability to separate official edits (by designated curators of the data and the metadata) from crowd-sourced content. The implementation allows a data curator, such as a statistical agency, to collect and incorporate improvements suggested by knowledgeable users in a structured way.