NSF-Census Research Network
Newsletter
Vol. 1, Issue 4

NCRN Fall Meeting held in NYC

Over 30 people attended the Fall NCRN meeting in New York City September 11-12.

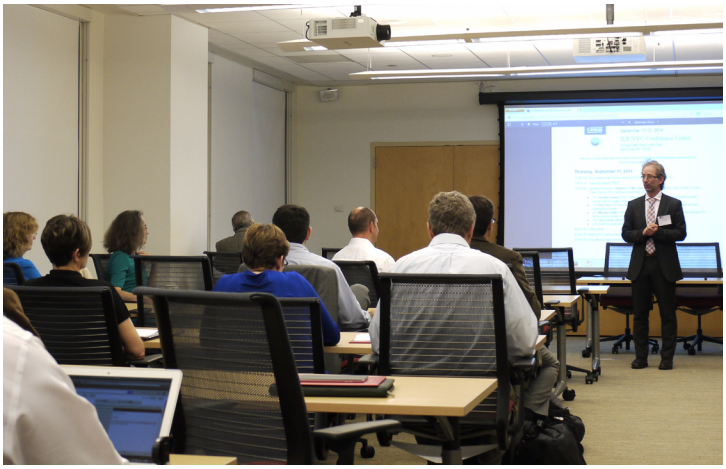
The first day's session was about Statistics of Non-standard Data, organized by Lars Vilhuber (Cornell node) and was chaired by Connie Citro (National Academy of Sciences). Presenters included: Jennifer Parker (CDC); Abe Dunn (BEA); Trey Spiller (CDC). Bill Eddy (CMU node), was the discussant. The presentations illustrated how statistical agencies can take advantage of administrative data and network sampling, and how they can leverage data from multiple sources to improve inferences and data products.

The following day's session focused on geo-spatial statistics. Scott Holan, (University of Missouri node), chaired and organized the session. Daniel Brown, (U. Michigan node), Rebecca Steorts, (CMU node); Jonathan Bradley, (U. Missouri node); Harrison Quick, (U. Missouri node) were the presenters. Kevin Konty, (New York City Department of Health and Mental Hygiene) was the discussant. The presentations featured research from multiple NCRN nodes, including improved methods for defining geographic regions for tabulations, for obtaining small area estimates, and for disseminating confidential data with point-referenced geographies. Kevin Konty noted that the techniques being developed by NCRN researchers are highly relevant for local and state government agencies, and provided numerous examples.

A link to see the fall meeting's presentations can be found on the NCRN website.

The next meeting will be held May 20-21 in Washington DC at the U.S. Census Bureau.

Jennifer Parker from the CDC addresses the participants of the NCRN Fall meeting in New York City.



Lars Vilhuber, Cornell Node, interacts with Fall NCRN meeting attendees.

Good discussions ensued even during break times.

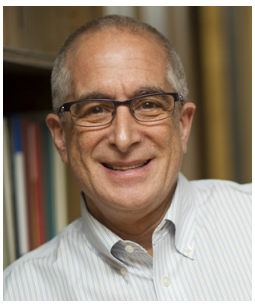
Follow Us on Twitter!

Have a Twitter account? You can follow NCRN on Twitter @NCRNCO.

Focus on Research Node: Cornell University

The NSF Census Research Network Cornell Node on "Integrated Research Support, Training and Data Documentation" started as a collaborative effort between researchers at three departments and centers at Cornell: the Labor Dynamics Institute (LDI) within the School of Industrial and Labor Relations (Department of Economics); the Department of Statistical Science; and the Cornell Institute for Social and Economic Research (CISER). The node combines researchers with long years of experience working with statistical agencies, in particular the Census Bureau, with new researchers bringing novel approaches to the table. The node's research foci include cutting-edge methods from computational statistics, applied to Census Bureau problems, and bringing a new approach to the dissemination and protection of metadata, and to metadata provenance. The node's researchers also continue work on new confidentiality protection methods for very large, complex and microdata structures, on the economics of disclosure avoidance, and on the creation of new statistics derived from linked employer-household data.

The team includes PI **John Abowd**, Edmund Ezra Day Professor of Economics, Director of the Labor Dynamics Institute and Professor of Statistics and Information Science; PI **Lars Vilhuber**, Executive Director of the Labor Dynamics Institute and Senior Research Associate in the ILR School (Department of Economics); PI **William Block**, Director of the Cornell Institute for Social and Economic Research and Executive Director of the Social Science History Association; and PI **Ping Li**, currently Associate Professor at the Department of Statistics and Biostatistics and the Department of Computer Science at Rutgers University. Senior researchers include **Warren Brown**, Senior Research Associate of the Cornell Institute for Social and Economic Research, and **Carl Lagoze**, who is Associate Professor at the School of Information at the University of Michigan. Both Li and Lagoze were at Cornell University at the start of the grant. The team also includes several staff members (all of whom are housed at CISER), and two doctoral students (one in statistics, one in economics).



NCRN Cornell PI

The researchers at Cornell University were working with U.S. Census data long before the NCRN project began, which made it a natural fit for them to apply for the grant. Abowd and Vilhuber were part of the founding team in the Census Bureau's Longitudinal Employer-Household Dynamics Program.

One of the node's software projects, captured under the umbrella of the Comprehensive Extensible Data Documentation and Access Repository (CED²AR), allows for metadata capture, enhancements to metadata standards, and simultaneous discovery and dissemination in public and restricted-access research environments. "When we started this project, there was surprisingly little work on how to protect confidential metadata

Around the table clockwise from bottom right: NCRN staff members Venkata Kambhampaty and Benjamin Perry, NCRN PI Lars Vilhuber (via video conference), NCRN Senior Researcher Warren Brown, and NCRN PI Bill Block.

(data about data), despite many, many years of research into, and solutions for, the protection of confidential data" noted Vilhuber. "The default solution – to completely avoid publishing metadata on the confidential data – leads to an incredible paucity of documentation about those confidential data" at statistical agencies such as the Census Bureau. To allow for the publication of more information about confidential research data, Cornell is working on enhancing a widely used metadata standard, the Data Documentation Initiative (DDI), however the principles are general and can be applied to a variety of other standards. "We started out enhancing the standard to provide fine-grained control about the release of information elements," remarked

"The default solution – to completely avoid publishing metadata on the confidential data – leads to an incredible paucity of documentation about those confidential data"

Vilhuber, "such as the name of a variable itself, or some of its statistical properties such as certain extreme values." The enhanced standard has since been proposed to the DDI organization for adoption. "But once we had the standard, we could not use any of the existing software to easily highlight and leverage our contribution, so we started to build our own software to edit and publish enhanced DDI." The resulting web software has been actively in use at their website for two key datasets, the Survey of Income and Program Participation (SIPP) Synthetic Beta and the

Continued on page 4

Upcoming NCRN Virtual Seminars

The NCRN Virtual Seminar is held at different locations throughout the network, every first Wednesday of the month. Nodes engage in the seminar using videoconferencing equipment, from on-campus locations as listed under each event. Active participants can ask questions through the live video feed.

Participation via videoconference is open to interested parties, in particular at other non-NCRN universities and statistical agencies at the federal, state, and local level, as long as equipment requirements can be satisfied. For more information, contact info@ncrn.info.

November 5

Speaker: **Nicholas Nagle (University of Tennessee at Knoxville)**

Title: *Survey weighting with informative but imprecise benchmarks*

Abstract: Survey weights are often adjusted so that the estimated totals match with known benchmark totals. This practice is limited by the requirement that benchmarks be perfectly known and the tendency for survey weight variability to increase as more benchmarks are included. We modify the Iterative Proportional Fitting adjustment method to incorporate benchmarks that are imprecisely known. This allows the use of benchmark controls from sources of data that are not currently eligible for benchmarking too, such as auxiliary surveys or other incomplete records. This method also allows us to efficiently increase the number and types of benchmark data that are used for survey weighting. We present results from efforts to adjust public use microdata samples to generate estimates and microsimulations for small areas (i.e. tracts and block groups).

December 3

Speaker: **Bruce D. Spencer and Zachary H. Seeskin (Northwestern University)**

Title: *Measuring Benefits from Improving Accuracy of the 2020 Census: Apportionment of the U.S. House of Representatives and Allocation of Federal Funds*

Abstract: In order to know how much accuracy is needed for the 2020 Census – with the appreciation that accuracy is expensive – we need to understand how the census results get used. In this talk, we consider two high profile uses of the census: apportionment and fund allocation. Apportionment of the 435 seats in the U.S. House of Representatives is based on census numbers, and distortions in the census results lead to distortions in numbers of seats allocated to the states. We expect that roughly $5 trillion in federal grant and direct assistance monies will be distributed at least partly on the basis of population and income data following the 2020 census, and distortions in census results cause distortions in the allocations of funds. We present loss functions to quantify the distortions in apportionment and fund allocations, and we describe empirical analyses to estimate the expected loss arising from alternative profiles of accuracy in state population numbers.

Other Events

The Federal Committee on Statistical Methodology (FCSM) Statistical Policy Seminar will be held December 15-16, 2014 at the Washington Convention Center, 801 Mount Vernon Place, Washington DC. Several presentations from NCRN nodes are expected to take place.

Cornell Node (Continued)

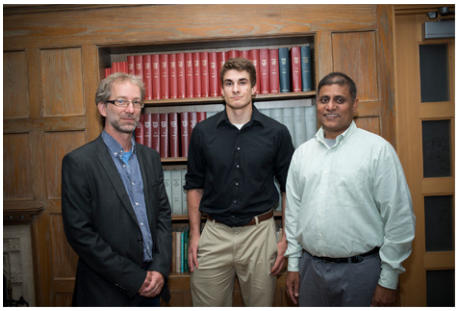
from page 2

Synthetic Longitudinal Business Database (SynLBD). Work is underway on a DDI-generating workflow customized for the Center for Economic Studies at the Census Bureau. Vilhuber also mentioned that the Cornell node is developing additional applications that specifically take into account the context of working within secure data enclaves, such as the Census Bureau's Research Data Centers, CISER's Cornell Restricted Access Data Center (CRADC) and LDI's Synthetic Data Server. "Most researchers are not data archivists, nor do they aim to be," Vilhuber added, "Software that captures the metadata that only researchers know must make their activities easier along some other dimension, not harder. Most metadata capture tools don't live up to that standard."

"When we started this project, there was surprisingly little work on how to protect confidential metadata (data about data)"

The other core activity at Cornell is the use of high-performance logistic regression methods as a tool for data edits and imputation. Some of the uses include multiple response variables, such as the race and ethnicity questions on the decennial census and on the American Community Survey, as well as incompletely coded links, for instance the location of establishments that employ workers in linked employer-household data (in the Census Bureau's LEHD data, only the employer, but not the establishment, is encoded for most states' data).

Besides being itself a widely dispersed node, with active members at the University of Michigan, Rutgers University and the University of California at Berkeley (where PI John Abowd is spending his sabbatical this year), the node has active collaborations with many other nodes. Some of this work is funded through the NCRN grants on either end, other work predates the NCRN grants, or was separately initiated. Cornell has collaborated for several years with members of the Duke University/NISS node (as well as Census Bureau staff) on the Survey of Income and Program Participation (SIPP) synthetic dataset (SIPP Synthetic Beta) and the Synthetic Longitudinal Business Database (SynLBD). As part of the CED²AR work described earlier, the node created new and enhanced online codebooks, naturally encoded with enhanced DDI, and served by CED²AR software to the public. Training sessions on the use of the SIPP Synthetic Beta, organized by the Michigan node and held at Duke and at Michigan, were supported by members of the Cornell node, and used the CED²AR-hosted codebook for the SIPP Synthetic Beta as a training tool.



NCRN Cornell PI Lars Vilhuber, and NCRN staff members Benjamin Perry and Venkata Kambhampaty.

"One of the advantages of the network is that we can broaden the reach we are getting beyond just the research nodes," noted Vilhuber, "Our scope of impact has been much larger...there is a much stronger network of collaboration and activity since the beginning of NCRN."

Save the Date!

NCRN Spring Meeting
May 20 - 21, 2015
Washington DC
Details in our next newsletter.

Node News

Krista Park New NCRN Rotational Staff Member

Krista Park has been selected as the Census Bureau's first 6-month NCRN rotational staff member. She has been with the Census Bureau since 2007 in the Geographic Support Branch in Field Division. She managed the division's activities for the Local Update of Census Addresses (LUCA) and Count Question Resolution (CQR) programs and served as the non-supervisory team lead for the Partnership Geography programs during the 2010 Census. More recently, Krista has led a focus group research team designing improvements for the LUCA program for the 2020 Census and provided support for frame improvement modeling projects for the 2020 Census.

Krista earned a Ph.D. in American Studies with a focus on skill-based communities from the University of Maryland. She received B.A.s in Literature and Music from American University and an M.A. in American Studies from the University of Wyoming. She also has a Masters Certificate in Project Management from George Washington University and maintains a Project Management Professional (PMP) Certification through the Project Management Institute.

Krista Park, NCRN Rotational Staff Member at the U.S. Census Bureau.

Acquisti Gives Symposium for Installation of 14th President of the University of Michigan

Alessandro Acquisti, professor of information technology and public policy at the Heinz College, Carnegie Mellon University (CMU), co-PI of NCRN CMU node, and co-director of the CMU Center for Behavioral and Decision Research, recently gave the invited inauguration symposium on the occasion of President Mark S. Schissel's installation as the 14th President of the University of Michigan on September 5. The topic was "Privacy and Identity in a Hyperconnected Society." Personal privacy concerns are becoming more of an issue, especially when looking at areas such as education, commerce, social interactions, policing, national security, health, and social research to name a few. Acquisti talked about results from several studies and experiments addressing the behavioral economics of privacy in online social networks. There are many trade-offs that emerge from the protection or sharing of personal information, the inadequacy of "notice and consent" mechanisms for privacy protection . He also talked about in the near future, an augmented-reality world in which online and offline personal data will seamlessly blend.

PI Role Shifts to Lars Vilhuber

Alan Karr, Director of the National Institute of Statistical Sciences (NISS) since 2000, and Principal Investigator on the NCRN Coordinating Office grant, left NISS at the end of August for his new position as Director of the Center of Excellence for Complex Data Analysis (CoDA) at RTI International. **Lars Vilhuber**, Executive Director of the Labor Dynamics Institute at Cornell University and co-PI on the NCRN Coordinating Office grant, has taken over as lead PI of the grant, while Karr will continue to serve as co-PI on the grant.

Publications

- Bradley, J.R., Holan, S.H., and Wikle, C.K. (2014) (Submitted) Mixed Effects Modeling for Aerial Data that Exhibit Multivariate-Spatio-Temporal Dependencies. arXiv:1407.7479
- Bradley, J.R., Wikle, C.K., and Holan, S.H. (2014) (Submitted) Bayesian Spatial Change of Support for Count-Valued Survey Data. arXiv:1405.7227
- Cox, L. (2014), Enabling statistical analysis of suppressed tabular data, in Privacy in Statistical Databases, edited by J. Domingo-Ferrer, Lecture Notes in Computer Science 8744. Heidelberg: Springer, 1-10.
- J. Hu, J. P. Reiter, Q. Wang, (2014), Disclosure risk evaluation for fully synthetic data, in Privacy in Statistical Databases, edited by J. Domingo-Ferrer, Lecture Notes in Computer Science 8744. Heidelberg: Springer, 185-199.
- Kim, H. J., Karr, A. F., and Reiter, J. P. (2014), Statistical Disclosure Limitation in the Presence of Edit Rules, Journal of Official Statistics, forthcoming
- Kim, H. and S. N. MacEachern (forthcoming), "The generalized multiset sampler," Journal of Computational and Graphical Statistics.
- D. Manrique-Vallier (forthcoming), "Longitudinal mixed membership trajectory models for disability survey data," Annals of Applied Statistics.
- McElroy, T.S. and Holan, S.H. (2014) Fast Estimation of Time Series with Multiple Long-Range Persistencies, *ASA Proceedings of the Joint Statistical Meetings, American Statistical Association (Alexandria, VA)*.
- Porter, A.T., Holan, S.H., and Wikle, C.K. (2014) (Submitted) Bayesian Semiparametric Hierarchical Empirical Likelihood Spatial Models. arXiv:1405.3880
- Porter, A.T. and Oleson, J. (2014) (To Appear; *Spatial and Spatio-Temporal Epidemiology*) A CAR Model for Multiple Outcomes on Mismatched Lattices.
- Porter, A. T., Holan, S.H., Wikle, C.K., and Cressie, N. (2014) (To Appear; *Spatial Statistics*) Spatial Fay-Herriot Models for Small Area Estimation with Functional Covariates. arXiv:1303.6668
- Quick, H., Holan, S.H., Wikle, C.K., and Reiter, J.P. (2014) (Submitted) Bayesian Marked Point Process Modeling for Generating Fully Synthetic Public Use Data with Point-Referenced Geography. arXiv:1407.7795
- J. P. Reiter (2014), "A case for public access to redacted social science data," FierceBigData, Sept. 3, 2014.
- Wikle, C.K. (2014) (To Appear) Modern Perspectives on Statistics for Spatio-Temporal Data. *WIRES Computational Statistics.*
- Wikle, C.K. (2014) (To Appear) Agent Based Models: Statistical Challenges and Opportunities. *Statistic Views,* Wiley.
- Yang, W.H., Holan, S.H., and Wikle, C.K. (2014) (Submitted) Bayesian Lattice Filters for Time-Varying Autoregression and Time-Frequency Analysis. arXiv:1408.2757
- Zhuang, L. and Cressie, N. (2014) (To Appear) Bayesian Hierarchical Statistical SIRS Models. *Statistical Methods and Applications.*

Presentations

- Acquisti, A. “Privacy and Identity in a Hyperconnected Society.” Inauguration of President Mark S. Schissel, University of Michigan, Ann Arbor, September 5, 2014 (see related article on page 5).
- Holan, S.H., Bayesian Dynamic Time-Frequency Estimation, Twelfth World Meeting of ISBA, Cancun, Mexico. July 2014.
- Reiter, J. “Generating and releasing synthetic data: Lessons learned and future directions,” MITRE workshop, McLean, VA, July 2014.
- Reiter, J. “Providing public access to confidential, big social science data,” ISNIE conference, Duke University, Durham, NC, June 2014.
- Quick, H., A Fully Bayesian Approach for Generating Synthetic Marks and Geographies for Confidential Data, International Indian Statistical Association (IISA), Riverside CA. July 2014.
- Quick, H., A Fully Bayesian Approach for Generating Synthetic Marks and Geographies for Confidential Data, Joint Statistical Meetings – New Researcher’s Conference, Boston, MA. August 2014.
- Wikle, C.K., Ecology of infectious disease, Invited Discussion, SAMSI Program on Mathematical and Statistical Ecology: Opening Workshop. August 2014.
- Wikle, C.K., Statistics for Spatio-Temporal Data Tutorial, Invited Tutorial Lecture, SAMSI Program on Mathematical and Statistical Ecology: Opening Workshop. August 2014.
- Wikle, C.K., Interaction-based parameterizations for nonlinear dynamic spatio-temporal statistical models. Keynote Lecture: Twelfth World Meeting of ISBA, Cancun, Mexico. July 2014.
- Wikle, C.K., Ecological prediction with high-frequency “big data” covariates. Plenary Lecture: International Statistical Ecology Conference, Montpellier, France. July 2014.