NSF-Census Research Network
Newsletter
Vol. 2, Issue 2

Research Node Focus: Northwestern University

Northwestern University's NCRN node is focusing on two major research areas. One, led by **Charles (Chuck) Manski,** Board of Trustees Professor in the Department of Economics, Faculty Fellow in the Institute for Policy Research, and Co-PI of the Northwestern University research node, concerns different sources of error in statistics. There are some types of errors in official statistics that the statistics community typically does not address in its approach to total survey error. An example of that is *revision error*. "So, for example, statistics might be released showing how fast the economy is growing, and then a quarter later, when additional data and metadata are available, the statistics are revised and could indicate quite a different rate of growth," explained **Bruce Spencer**, Professor of Statistics, Faculty Fellow of the Institute of Policy Research, and PI of the Northwestern University research node. The revisions can be quite substantial, so revision reflects an important source of uncertainty that often is neglected. The probable or historical extent of revision needs to be reported as a source of uncertainty when the statistics are released to policy makers.

"Some people don't want to deal with uncertainty, but we in the profession could be better at reporting our data in a way that recognizes the uncertainty." said Spencer. For example, if one were to be reporting the unemployment rate, instead of saying the unemployment rate is "X", one would say, "the unemployment rate is estimated to be 'X' with a margin of error of +/- 5 percent."

"When we don't report the uncertainty level, people more easily become overconfident about what the statistics are," Spencer remarked.

Manski is also studying how to improve statistical inference in the presence of missing data, and is developing error bounds for different components of error that either do not depend on the statistician's judgment or that do so in transparent ways. Often statisticians use some judgment when estimating the effect of missing data or non-response. For example, often the judgment asserts that the data are missing at random, so that once the statistician conditions on some covariates, there is no bias from the missing responses. Manski has been developing error bounds to account for missing data that do not depend on the statistician's discernment or that do so in transparent ways. "When we are using our judgments, we need to recognize that, usually, our judgment is to some degree wrong. Manski considers the range of values for a statistic that are credible when no expert judgment is invoked but only the empirical data. As you bring in some simple assumptions, how does that range narrow? That allows you to see how much the reduction in the uncertainty is due to the making of assumptions. When the assumptions have large effects on the intervals, one should thoughtfully consider the uncertainty of those assumptions, because ignoring that uncertainty leads to overconfidence. So a lot of this has to do with making the statistics more transparent," Spencer commented.

Northwestern University. (Photo courtesy of Northwestern University Media Relations Office)

Spencer's work revolves around the value of statistics. The value of statistics arises because the particular statistics are being used. The notion is that if the statistics are more

Continued on page 2

In This Issue

Northwestern University Node (Continued)

accurate or of higher quality, then the uses are better. One area Spencer has looked at is uses for allocating funds and for apportioning representation in Congress. Apportionment is the key reason for there being a constitutional requirement for having a census. Zach Seeskin, a graduate student in the Department of Statistics and Institute for Policy Research, is conducting part of his doctoral dissertation research on this question. “We are looking at the effects of alternative levels of error in the 2020 census,” said Spencer. They are working with the Census Bureau to do a cost-benefit analysis of alternative 2020 census designs. They are specifically looking at ways to compare the need for better data for allocating funds and allocating representation to the cost of getting better data. “One of the areas that matters most is errors in state population numbers. This is not true for all matters, but is still of importance. Ethnicity, age, gender and other factors are also important. But for apportioning money, for example, it could be $500 billion over a 10-year period, the state population number is extremely important. How does the distribution of seats in the U.S. House of Representatives change when the accuracy of the U.S. Census changes?” explained Spencer.

They are looking at the formula that is used to divide up the seats in the U.S. House and they are examining how many of these seats are going to the wrong states for different levels of accuracy. The various choices will be presented, but no one scenario will be recommended, as Spencer noted that would be a political judgment. They will list the various options along with their accuracy and effects on apportionment and allocation and the associated census cost.

For the fund allocation, Spencer and Seeskin started with an identified set of formulas for grants in aid programs (or legislation for intergovernmental transfers) that use population data. They selected a carefully designed sample of those programs and they have been studying what is the effect of different census accuracy on how much money goes to the wrong states over a ten-year period. Spencer and Seeskin are drafting a paper that they will be sharing with the U.S. Census Bureau and, once they get the Bureau’s comments and feedback, they will put the paper out for publication.

The other area of research that Spencer and Seeskin are working on is how official statistics get used in decision making and policy making. Spencer said, “Uses of statistics for policy making is potentially very important, but is also very difficult to evaluate.” We want to go beyond saying someone used statistics to make a decision; what we want to know, is, if the decision maker didn’t have the statistic available, or the statistic was at a different level of accuracy, would the same decision have been made or a different decision? “It is a counter-factual. If we had different data how would the outcomes have been different? But we can only observe what happened with official statistics as we have them and not what would have happened with different statistics available. So we are having to make causal inferences about the effects of additional statistical measures or changes in the statistical quality,” said Spencer.

It is even hard to identify whom the decision makers are that one would ask how he or she would make a different decision if more accurate data was available.

Seeskin plans to continue working on this type of research after he completes graduate school. He is also doing additional research, working with the Census Bureau on using big data to improve imputations in the American Community Survey.

Spencer explained that the other Northwestern statistics students are also learning that this type of research is an important area of research. “I’ve also incorporated some additional material about official statistics in my course for first year graduate students in statistics, so they can see some of the problems we are working on and some of the other NCRN nodes are working on,” said Spencer.

Spencer remarked, “One of the other advantages is we are really applying statistical decision theory to the problems of governmental statistical agencies. Over the decades there has been less and less teaching of applied statistical decision theory. So, many graduate students might not be exposed to it. It’s a way of helping to further develop an important tool and keep the methodology going.”

Call for Papers Now Open for FCSM Research Conference

On December 1-3, 2015, over 800 active, innovative people from government, academia, and business will convene at the Washington, DC Convention Center to learn and share at the Federal Committee on Statistical Methodology (FCSM) Research Conference. The three day conference provides a forum for experts from around the world to discuss and exchange current research and methodological knowledge relevant to statistical programs.

Papers are invited on a wide range of topics and must be original and not previously published or disseminated. They may be submitted as individual papers or as part of a broad-based organized session.

LATE BREAKING News: NCRN Meetings were held at the National Academies of Science and the U.S. Census Bureau

On May 7 and 8, over 300 participants from a wide range of statistical agencies, academic institutions and research institutes heard presentations from U.S. Census Bureau and NCRN researchers.

On May 7, the Principal Investigators of the NCRN nodes met with Census Bureau collaborators, including Director John Thompson and Associate Director Nancy Potok. Three parallel sessions had participants see the fruit of collaborations between NCRN nodes and the Census Bureau, as well as new research challenges faced by statistical agencies.

On May 8, NCRN node members gave 13 presentations on topics such as the training of the next generation of methodologists, uses and benefits of government statistics, geographic aspects of statistics, confidentiality, and the statistics of unstructured data. Discussants from statistical agencies, academia, and government provided lively and useful feedback, as did the attending audience. The highlight of May 8 was the Public Seminar of the 127th CNSTAT Meeting, with NCRN PIs John Abowd and Stephen Fienberg presenting their outlook on "Can Government-Academic Partnerships Help Secure the Future of the Federal Statistical System? Examples from the NSF-Census Research Network", and panelists Robert Groves (Georgetown University and former Director of the U.S. Census Bureau) and Erica Groshen (Commissioner, Bureau of Labor Statistics) providing further discussion.

The next NCRN Meetings are currently being planned, please consult ncrn.info for any updates.

Node News

Laura Brandimarte, postdoctoral fellow in the Heinz College, Carnegie Mellon University, has accepted a position as assistant professor in the Department of Management Information Systems of the Eller College of Management, University of Arizona, Tucson, AZ.

Stephen E. Fienberg will deliver the COPSS R.A. Fisher Lecture, "R.A. Fisher and the Statistical ABCs," at the 2015 Joint Statistics Meetings in Seattle, WA in August.

Mauricio Sadinle, Department of Statistics, Carnegie Mellon University, successfully completed his Ph.D. dissertation, "A Bayesian Partitioning Approach to Duplicate Detection and Record Linkage,“ in January 2015, and has joined the Duke/NISS node as a postdoctoral fellow.

Mauricio Sadinle.

Rebecca Steorts, Visiting Assistant Professor in the Department of Statistics, Carnegie Mellon University, has accepted a position as assistant professor in the Department of Statistical Science, Duke University. She will also be joining the Duke/NISS NCRN node.

Steorts also was recently awarded a grant from Knowledge Lab at the University of Chicago via the John Templeton Foundation. Rebecca is one of the newest member's of the Knowledge Lab's Metaknowledge Research Network. Her research will involve Bayesian models to cluster records together to link records and deal with duplicated information with applications to human rights violence, medical applications, official statistics, among others. Steorts' work will scale to large and complex data sets by adding innovative blocking methods from computer science for limiting the search space of her algorithms. Steorts plans to make her research available as open-source software packages, and can be used by fellow NCRN researchers amongst those in social science, computer science, machine learning, and statistics.

Rebecca Steorts.

Hang Kim, postdoctoral fellow at the National Institute of Statistical Sciences and Duke University, has accepted a position at the University of Cincinnati as an assistant professor in the Department of Statistics.

The Cornell NCRN node welcomes **Long Zhang** as a postdoctoral fellow. Long is a statistician and data scientist who works on machine learning and statistical disclosure limitation. He comes to Cornell from the University of Florida, Gainesville, Department of Statistics where he completed his Ph.D. in December 2014. He will work with NCRN PIs **John Abowd** and **Lars Vilhuber** on two NSF-funded projects (TC: Large and NCRN-Cornell), working on the interface between economics, computer science, and statistics to develop disclosure avoidance methods appropriate for complex economic microdata. Long started on April 1, 2015.

Noel Cressie's (Missouri Node) book, *Statistics for Spatial Data, revised edition,* Wiley, New York, NY, 1993 (900 pp.), has been inducted into the Wiley Classics Library in 2015.

Cressie has been named as a Member, ASA Advisory Committee on Climate Change, 2011-2017. He is also a member, COPSS Fisher Award Selection Committee this year. He was on the Scientific Committee, The International Environmetrics Society (TIES) for the 2014 Meeting, held in Guangzhou, China, December 2014.

Cressie was also the co-organizer and Chairman, Invited Paper Session on "Statistical Inference with Dependent Data," ISI Regional Statistics Conference, Kuala Lumpur, Malaysia, November 2014.

A new book entitled *Handbook of Discrete-Valued Time Series*, will be published by Chapman and Hall/CRC in 2015. The book was edited by Richard A. Davis, **Scott H. Holan** (Missouri), Robert Lund, and Nalini Ravishanker.

Rebecca Nugent (Carnegie Mellon) is this year's winner of the American Statistical Association Waller Education Award, which honors individuals for innovation in Statistics instruction.

Publications

The following are the most recent additions to publications produced by the research nodes within NCRN. A comprehensive list can be found here. (http://www.ncrn.info/documents/bibliographies)

Acquisti, Alessandro, Laura Brandimarte, and George Loewenstein. "**Privacy and human behavior in the age of information**." *Science* 347, no. 6221 (2015).

Bradley, J. R., C.K. Wikle, and S. H. Holan. "**Bayesian Spatial Change of Support for Count–Valued Survey Data**." *ArXiv e-prints* (2015).

Bradley, J.R., C.K. Wikle, and S.H. Holan. **Multiscale Analysis of Survey Data: Recent Developments and Exciting Prospects**, *Statistics Views*., 2015.

Bradley, J. R., C.K. Wikle, and S. H. Holan. "**Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics**." *ArXiv e-prints* (2015).

Cressie, N., S. Burden, W. Davis, P. Krivitsky, P. Mokhtarian, T. Seusse, and A. Zammit-Mangion. "**Capturing multivariate spatial dependence: Model, estimate, and then predict**." *Statistical Science* (2015).

Cressie, N., and R. L. Chambers. "**Comment: Spatial sampling designs depend as much on "how much?" and "why?" as on "where?"**." *Bayesian Analysis* (2015). A comment on "Optimal design in geostatistics under preferential sampling" by G. da Silva Ferreira and D. Gamerman.

Folch, D., and S. E. Spielman. "**Reducing the Margins of Error in the American Community Survey Through Data-Driven Regionalization**." *PlosOne* (2015).

Holan, S.H., and C.K. Wikle. "**Hierarchical Dynamic Generalized Linear Mixed Models for Discrete--Valued Spatio-Temporal Data**." In *Handbook of Discrete--Valued Time Series*., 2015.

Lund, R., S.H. Holan, and J. Livsey. "**Long Memory Discrete--Valued Time Series**." In *Handbook of Discrete-Valued Time Series*. Chapman and Hall, 2015.

Porter, A.T., C.K. Wikle, and S.H. Holan. "**Small Area Estimation via Multivariate Fay-Herriot Models With Latent Spatial Dependence**." *Australian & New Zealand Journal of Statistics* 57 (2015): 15-29.

Quick, H., S. H. Holan, C. K. Wikle, and J. P. Reiter. "**Bayesian Marked Point Process Modeling for Generating Fully Synthetic Public Use Data with Point-Referenced Geography**." *ArXiv e-prints* (2015).

Schifeling, T., C. Cheng, D. S. Hillygus, and J. P. Reiter. "**Accounting for nonignorable unit nonresponse and attrition in panel studies with refreshment samples**." *Journal of Survey Statistics and Methodology* (forthcoming).

Wikle, C.K., and S.H. Holan. "**Comment on ``Semiparametric Bayesian Density Estimation with Disparate Data Sources: A Meta-Analysis of Global Childhood Undernutrition" by Finncane, M. M., Paciorek, C. J., Stevens, G. A., and Ezzati, M.**" *Journal of the American Statistical Association* (2015).

Wikle, C.K., and M.B. Hooten. "**Hierarchical Agent-based Spatio-temporal Dynamic Models for Discrete Valued Data.**" In *Handbook of Discrete--Valued Time Series*., 2015.

Yang, W. H., S. H. Holan, and C.K. Wikle. "**Bayesian Lattice Filters for Time-Varying Autoregression and Time-Frequency Analysis**." *ArXiv e-prints* (2015).

Zhuang, L., and N. Cressie. "**Bayesian Hierarchical Statistical SIRS Models**." *Statistical Methods and Applications* 23 (2015): 601-646.

Presentations

December 2, **Ben Perry** gave a presentation on "Collaborative Editing and Versioning of DDI Metadata: The Latest from Cornell's NCRN CED²AR Software" at the 6th Annual European DDI User Conference in London. View the full program here: http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi14/schedConf/program

Noel Cressie, (Missouri) was the keynote speaker at the Geosummit in Redlands, California January 22. His talk was entitled, "Spatio-temporal statistics in geodesign."

On January 30, **Chris Wikle** (Missouri) presented a talk, "Quantifying Spatial Aggregation Error Using Weighted Eigenfunctions" at the Texas A&M Spatial Statistics workshop.

On March 19, **Isaac Sorkin** (Michigan) gave a talk at the U.S. Census Bureau, "Ranking Firms Using Revealed Preference" in the CES Seminar.

Scott Holan (Missouri) presented "Soil Property Estimation and Design for Agroecosystem Management using Hierarchical Geospatial Functional Data Models" at the Environmental Protection Agency in Washington, D.C. on March 25.

Holan also lectured at the University of Illinois, Urbana-Champaign on April 2. His talk was "Bayesian Lattice Filters for Time-Varying Autoregression and Time-Frequency Analysis."

Jerry Reiter (Duke) gave a talk: "Protecting Data Confidentiality in an Era with No Privacy" at the National Institute of Health in Washington D.C. on April 10.

Allan McCutcheon (Nebraska) was invited to give the 2015 Joint Program in Survey Methodology (JPSM) Distinguished Lecture on April 10 at the University of Maryland. He spoke on "Web Surveys, Online Panels, and Paradata: Automating Responsive Design." Recent distinguished lecturers include Mike Traugott, Norbert Schwarz, and J.N.K. Rao. You can see all of the past lecturers here.

Duke University held a workshop on the Redesign of the Survey of Income and Program Participation April 25 in San Diego, California. This workshop, which is part of the Population Association of America conference, provided introduction, background, and context of the SIPP's current and past designs for new and current SIPP researchers. The workshop gave an overview of SIPP's content, file structure, and data availability. It demonstrated some of the possible ways to access and use SIPP data and it provided an opportunity to increase stakeholder involvement and interaction with the SIPP program staff.

The NCRN Newsletter is published quarterly by the NCRN Coordinating Office. You may reach us at:
275 Ives Hall,
Cornell University, Ithaca, NY 14853
(607) 255-2744
info@ncrn.info
www.ncrn.info
Jamie Nunnelly - Managing Editor