NSF-Census Research Network
Newsletter
Vol. 2, Issue 1

NCRN Spring Meeting in Washington DC

The NCRN will hold its Spring 2015 meeting in Washington D.C. In collaboration with the National Academies' Committee on National Statistics, **John M. Abowd** (Cornell Node), and **Stephen E. Fienberg** (Carnegie Mellon Node) will present at the prestigious semi-annual "CNSTAT Public Seminar," at the National Academies' headquarters on May 8, 2015. Their talk is tentatively titled "Can Government-Academic Partnerships Help Secure the Future of the Federal Statistical System?" In addition, node members will present talks in a variety of parallel workshops on themes (yet to be determined) in the morning, also at the National Academies.

On May 7, 2015, node members will be at the U.S. Census Bureau, for a number of research presentations with and for Census personnel, and to meet with their Census Bureau collaborators. More information will come available in early Spring, and will be made available at http://www.ncrn.info/event/ncrn-meeting-spring-2015.

Profile: University of Colorado-Boulder and University of Tennessee NCRN Research Node

University of Colorado-Boulder Assistant Professor of Geography **Seth Spielman** once owned a book store in the Flatiron Building in New York City, but now he is changing the map (literally) in the ways we look at Census data. He and **Nicholas Nagle**, He and Nicholas Nagle, the Principal Investigators of the joint University of Colorado-Boulder and University of Tennessee node of the NSF-Census Research Network (NCRN) are working on some very exciting research projects involving maps and small area estimates.

"The first thing we are trying, is to develop a case that geography and small area estimation are actually are not separate things," Spielman said.

Spielman noted that, people who trained to design design maps are rarely trained to design surveys (and vice versa), he believes that there is great potential in collaboration between these two groups. The group at Colorado is hoping to make the case that the design of geographies should be seen as part of the survey process. There are certain kinds of areas that the Census Bureau makes which are areas that are truly just for the purpose of statistical calculation. There are about 75,000 Census tracts in the USA, people often think about them as neighborhoods.

"Much like a jigsaw puzzle, the U.S. gets cut up into these 75,000 pieces. For each one of those jigsaw puzzle pieces, the Census Bureau produces an estimate of income, education and all of the other demographics. We are saying that how you make the jigsaw puzzle affects the quality of the data. People have concerns about data quality currently and one way to address that is to actually think about the jigsaw puzzle, not the survey design, but the actual pieces of the puzzle." noted Spielman.

Seth Spielman from University of Colorado-Boulder.

Another analogy Spielman made was to think about the data as having two jars of coins, one having all pennies in it, the other having a mix of many different coins. If you pull out three or four coins from the jar of pennies, you can quickly make an estimate of what is contained in the jar, while the same sample size from a jar of mixed coins is not

Continued on page 2

In This Issue

Colorado/Tennessee Node (Continued)

going to give you a good picture of what the contents are in the rest of the jar. "What we are asking is can you actually draw lines on a map that will look like a jar of pennies than a jar that is completely mixed. So you might have a rich neighborhood next to a poor neighborhood, so you could draw the lines where the neighborhoods are mixed together in a way that maybe they shouldn't be, because that would give you more heterogeneity and that might not let you figure out what is happening on the ground. Or you could draw the lines so the rich neighborhood is in one tract and the poor neighborhood is in another tract. We have designed computational tools that will make this easier to do," explained Spielman.

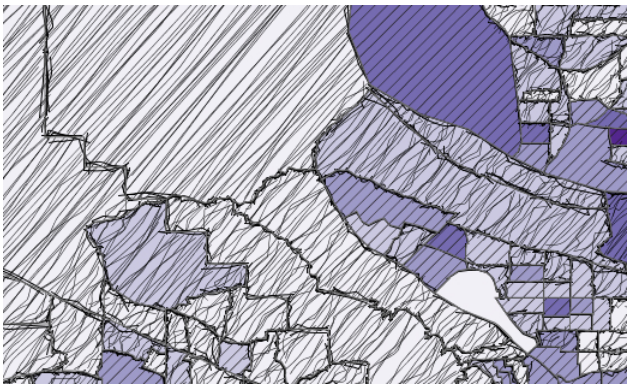
Spielman worked with sociologist John Logan and made maps of every single household in about 40 cities. They took some archival Census material and mapped cities at the individual level and that gave a very different picture of what the place looked like. "You could see an emergent structure in cities, so you see all these different dots in a city and you could find real neighborhoods and different kinds of neighborhoods just by looking at the people. The neighborhoods you see in cities are not simply a function of that jigsaw puzzle that the Census Bureau plops down, that there was something sort of underneath it that you don't normally see that had a natural organization to it," he said. He is working on using the Census Bureau master address file to apply those same types of insights to real cities and real places. "The thinking was that by using the geographic information to contain within survey responses we might be able to make better estimates. You would have better data without increasing the sample size and gain some insight into the underlying organization of populations. There was a double payoff of real sociological and geographic research about how people live and where they live and real benefits to data users as a form of higher quality data estimates," explained Spielman.

Nagle said his department is thinking about the margin of error that the American Community Survey (ACS) has in its small area estimations. The ACS has information about age by income, for example, and it has race by income, and race by home ownership, but it does not have all three of these items together. In other words, it would be hard to find out if African American households have a certain income level when they become home owners.

"We are designing new methods for integrating public use microdata and the Census tract data. We can produce estimates of these sorts of things. We can produce spatial microsimulations - to simulate households at Census tracts and so we could create custom tabulations of these households and estimate things like race, homeownership, and income," said Nagle.

Nagle is using techniques from the machine learning community because they also work with very large data sets and have similar problems with data quality. Machine learning has developed clever computational and statistical methods for addressing these problems and for controlling the uncertainty of the estimates - methods which transition nicely into solving our problems of estimating and simulating populations.

Nagle said the department is hiring a postdoc this year who will work on NCRN-related research. This person will look at how to improve the ACS by controlling the Census to more data. "One way to do this is to constrain the ACS to match up with data that we have from elsewhere," explained Nagle.



This map conveys uncertainty in estimates. The squigglier the lines, the more uncertain the data for each area of the map. This is one of many experimental maps the team produced in collaboration with Amy Griffin, a cartographer, who will present a virtual seminar in February (See page 3).

The other main activity that both Spielman and Nagle are working on is visualization. The group is looking at how to make maps that would represent the uncertainty or quality of the data. For example, if you have an estimated average home value of $200,000 but the margin of error is $150,000, how do you add that margin of error to the map?

Spielman said, "We have a web-based experimental protocol which involves eye tracking devices and all sorts of things. So we would ask the subject to identify the poorest section of the city, but each person would be shown a different map. Some maps might include just the poverty rate, some might include the poverty rate and the margin of error, some maps might include those things visualized in a different way. And so we look at how visualizing the quality of data affects the outcome in decisions."

So you may be wondering how Seth got to be the owner of an antiquarian bookstore? The store was actually first owned by his great aunt and uncle. One day he was visiting, helping them move boxes and at the end of the day, they asked him if he wanted to take over the store and he agreed. He owned the store for about five years, but it was during the time that Amazon started to take off, and he also started graduate school at Columbia. The rest was fate!

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.

For more information, go to: www.ncrn.info/events/virtual-seminar.

February 4

Speaker: Amy L. Griffin, University of New South Wales Canberra

Title: *Visualizing Attribute Uncertainty in the ACS: An Empirical Study of Decision-Making with Urban Planners*

http://www.ncrn.info/event/ncrn-virtual-seminar-feb-4-2015

Abstract: Recent changes to the U.S. Census have led to more timely updates of demographic statistics that are used in the delivery and planning of many social and environmental programs. However, this timeliness has a tradeoff: increased uncertainty in the estimates for small area geographies such as census blocks and tracts. Although the Census Bureau publishes information about the uncertainty of the estimates, few end users engage with and utilize this information, perhaps because it comes in a difficult to use form; another column in a table with many columns. Many techniques for visualising uncertainty in attribute data have been proposed, but few have been empirically tested, and fewer still with real end users using an ecologically valid task. Here, we report on a broader research program directed to studying the visualisation of attribute uncertainty for ACS data, and report the results of an experiment undertaken with 55 urban planners in which they had to make spatial decisions using uncertain demographic estimates. We compared visualisation methods based on two metaphors for communicating uncertainty: the stoplight and sketchiness. The experimental task is one taken from a context of use study we conducted on urban planning. It required planners to define an area of contiguous census tracts that meets a particular threshold with respect to the attribute in question: percentage of households in poverty. We conclude with some thoughts about how to help urban planners work with uncertainty in ACS data more effectively. (joint work with Jason Jurjevich, Portland State University, Meg Merrick, Portland State University, **Seth E Spielman**, Colorado University at Boulder, **Nicholas N Nagle,** University of Tennessee-Knoxville, David C Folch, Florida State University)

March 4, 2015

Speaker: John Abowd, Cornell University

Title: *Revisiting the Economics of Privacy: Population Statistics and Privacy as Public Goods*

http://www.ncrn.info/event/ncrn-virtual-seminar-march-4-2015

Abstract: We consider the problem of the public release of statistical information about a population–explicitly accounting for the public-good properties of both data accuracy and privacy loss. We first consider the implications of adding the public-good component to recently published models of private data publication under differential privacy guarantees using a Vickery-Clark-Groves mechanism and a Lindahl mechanism. We show that data quality will be inefficiently undersupplied. Next, we develop a standard social planner's problem using the technology set implied by (ε, δ)-differential privacy with (α, β)-accuracy for the Private Multiplicative Weights query release mechanism to study the properties of optimal provision of data accuracy and privacy loss when both are public goods. Using the production possibilities frontier implied by this technology, explicitly parameterized interdependent preferences, and the social welfare function, we display properties of the solution to the social planner's problem. Our results directly quantify the optimal choice of data accuracy and privacy loss as functions of the technology and preference parameters. Some of these properties can be quantified using population statistics on marginal preferences and correlations between income, data accuracy preferences, and privacy loss preferences that are available from survey data. Our results show that government data custodians should publish more accurate statistics with weaker privacy guarantees than would occur with purely private data publishing. Our statistical results using the General Social Survey and the Cornell National Social Survey indicate

Continued on page 4

Virtual Seminars (Continued)

that the welfare losses from under-providing data accuracy while over-providing privacy protection can be substantial.

April 1, 2015

Speaker: **Bimal Sinha**, University of Maryland, Baltimore County and **Marlow Lemons**, U.S. Census Bureau

Titles: "Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-normal Regression Samples" and "A Method to Improve Data Swapping at the U.S. Census Bureau"

http://www.ncrn.info/event/ncrn-virtual-seminar-april-2-2014

Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-normal Regression Samples (Bimal Sinha)

Abstract: In this article multiplication of original data values by random noise is suggested as a disclosure control strategy when only the top part of the data is sensitive, as is often the case with income data. The proposed method can serve as an alternative to top coding which is a standard method in this context. Because the log-normal distribution usually fits income data well, the present investigation focuses exclusively on the log-normal. It is assumed that the log-scale mean of the sensitive variable is described by a linear regression on a set of non-sensitive covariates, and we show how a data user can draw valid inference on the parameters of the regression. An appealing feature of noise multiplication is the presence of an explicit tuning mechanism, namely, the noise generating distribution. By appropriately choosing this distribution, one can control the accuracy of inferences and the level of disclosure protection desired in the released data. Usually, more information is retained on the top part of the data under noise multiplication than under top coding. Likelihood based analysis is developed when only the large values in the data set are noise multiplied, under the assumption that the original data form a sample from a log-normal distribution. In this scenario, data analysis methods are developed under two types of data releases: (I) each released value includes an indicator of whether or not it has been noise multiplied, and (II) no such indicator is provided. A simulation study is carried out to assess the accuracy of inference for some parameters of interest. Since top coding and synthetic data methods are already available as disclosure control strategies for extreme values, some comparisons with the proposed method are made through a simulation study. The results are illustrated with a data analysis example based on 2000 U.S. Current Population Survey data. Furthermore, a disclosure risk evaluation of the proposed methodology is presented in the context of the Current Population Survey data example, and the disclosure risk of the proposed noise multiplication method is compared with the disclosure risk of synthetic data.

A Method to Improve Data Swapping at the U.S. Census Bureau (Marlow Lemons)

Abstract: Data swapping is one of several disclosure avoidance methods that the Census Bureau implements to uphold confidentiality mandated by law. The Center for Disclosure Avoidance Research (CDAR) is currently studying the use of n-cycle swapping as a means to protect respondent identity in large-scale data. N-cycle swapping, a variant of data swapping, uses permutations of size 'n' to swap data records rather than swapping them in pairs. In this talk, we will discuss the processes surrounding n-cycle swapping, CDAR's current studies and challenges, and future projects and data products involving this disclosure avoidance technique.

Node News

Cornell University PI **John Abowd** has been elected as a new Fellow of the Econometric Society. The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics.

The work of **Seth Spielman** and **Nicholas Nagle** (see related article on page 1) was noted in a recent article in *City Lab*, a publication from *The Atlantic* magazine. The article, entitled "How to Make a Better Map - Using Neuroscience," (http://goo.gl/NJsSRV) referred to the Colorado/Tennessee node's research on how to better reflect uncertainty in the American Community Survey's (ACS) maps.

Noel Cressie, Missouri node, has been reappointed as a member of the U.S. Census Bureau's Census Scientific Advisory Committee (CSAC). His term will run from 2014-2017.

Krista Park, NCRN liaison at the U.S. Census, has been selected for a promotion and will be transitioning out as the NCRN liaison sooner than expected. She worked half time for NCRN in December and concluded her position on January 1, 2015. Nancy Bates, U.S. Census, said, "We greatly appreciate the inroads Krista has made during her short tenure as our liaison." A new liaison, **Renee Ellis**, has been announced (see related article below).

Census Bureau's NCRN Liaison Visits Research Nodes

During visits to three NCRN nodes in October 2014, Krista Park, the NCRN Liaison at the Census Bureau, presented information on hiring, fellowship, postdoc, and dissertation fellowship programs at the Census Bureau. In presentations at the University of Michigan (October 23, 2014), University of Nebraska – Lincoln (October 28, 2014), and at the University of Missouri (October 30-31), she spoke approximately 50 economists, statisticians, computer scientists, and students at the University of Missiouri's Truman School of Public Affairs.

On December 3, Krista presented the same information in a video teleconference before the monthly NCRN Virtual Seminar. Seventeen students from Carnegie Mellon University, Cornell University, Northwestern University, and the University of Michigan attended the session, drawn from economics, organizational behavior, policy analysis, regional planning, sociology, and statistics.

The PowerPoint slides for the presentation, which include URLs for additional information on the federal hiring process, are posted with the December 3, 2014, NCRN event information at http://www.ncrn.info/event/park-presentation.

NCRN Nodes Welcome Renee Ellis as the New NCRN Liaison at the Census Bureau

Renee Ellis is the new NCRN liaison at the U.S. Census Bureau. She replaces Krista Park, who was promoted to a new position at the Census Bureau.

Ellis joined the Census Bureau in 2010 in the Fertility and Family Statistics Branch in The Demographic Directorate. She works on various aspects of the redesign of the demographic and marital history sections for the Survey of Income and Program Participation including question design, instrument testing, edit specifications, and analysis.

She coordinated census contributions and served on the development and writing committees for the interagency children's forum that publishes the annual America's Children report as well as coordinating census data efforts for President Obama's "Brother's Keeper Initiative." Additionally, she has coauthored four census reports using SIPP, CPS and ACS data.

Ellis earned a Ph.D. in Sociology with a focus on demography, family, and the life course from the University of California Irvine. She received B.A.s in sociology and criminology from Weber State University and an M.A. in Sociology and Demography from the University of California, Irvine. She currently teaches demography and research methods as an adjunct professor at the University of Maryland University College.

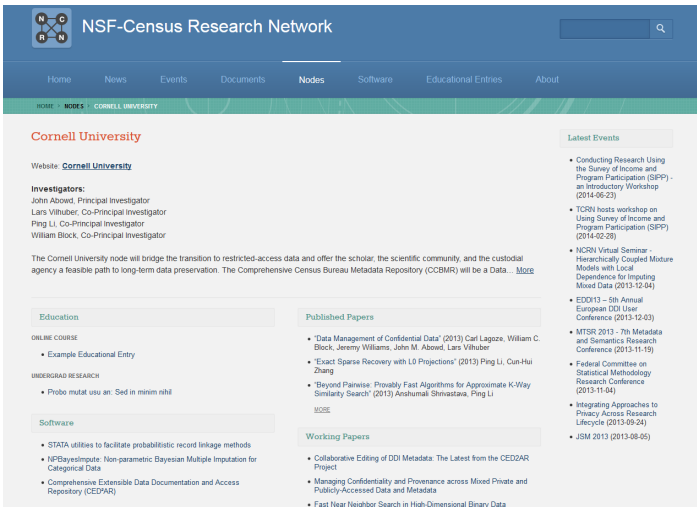
NCRN Website Has New Features

The NCRN website, www.ncrn.info, has a number of new features designed to surface more information on the nodes. Software that was developed and made available by the NCRN nodes are listed in a dedicated section. Educational activities are also now features in their own section. Both new sections can be filtered by node, or list all available entries network-wide.

Finally, the node profiles have been updated to reflect these new components, and provide a global view of all of the activities that a node is involved in: events, published papers as well as the most recent working papers and pre-prints, and, of course, educational activities and software. For instance, the profile for this issue's featured node at the University of Colorado at Boulder/University of Tennessee can be found at ncrn.info/node/11. The profiles for all nodes are listed at ncrn.info/nodes.

The bibliography section has been updated to show abstracts for papers, if available. The Coordinating Office staff will update all bibliographic entries with abstracts over the next coming weeks.

Node members wishing to add information on new publications, events, and node activities should contact info@ncrn.info with the details.



As always, interested parties can follow our RSS, Twitter (links listed on the home page) and subscribe to the calendar feed for activities such as the NCRN Virtual Seminars.

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)

Carrig, M. M., D. Manrique-Vallier, K. Ranby, J. P. Reiter, and R. Hoyle. (forthcoming), "A nonparametric, multiple imputation-based method for the retrospective integration of data sets," Multivariate Behavioral Research.

Cressie, N. and Burden, S. (2015). "Evaluation of diagnostics for hierarchical spatial statistical models," in K.V. Mardia Festschrift, eds I.L. Dryden and J.T. Kent. Wiley, Chichester, UK, forthcoming.

Folch, D. C., and S. E. Spielman. "The initialization of seeded regionalization algorithms." Journal of Geographical Systems (Submitted).

Folch, D., and S. E. Spielman. "Identifying regions based on flexible user defined constraints." International Journal of Geographic Information Science 28 (2014): 164-184.

Griffin, A. L., S. E. Spielman, J. Jurjevich, M. Merrick, N. N. Folch Nagle, and D. C. "Supporting planners' work with uncertain demographic data." In GIScience Workshop on Uncertainty Visualization. Vol. 23., 2014.

Griffin, A. L., S. E. Spielman, N. N. Nagle, J. Jurjevich, M. Merrick, and D. C. Folch. "Supporting planners' work with uncertain demographic data." In Proceedings of IEEE VIS 2014, 9-14. Proceedings of IEEE VIS 2014, 2014.

Manrique-Vallier, D., and J.P. Reiter. "Bayesian estimation of discrete multivariate latent structure models with structural zeros." Journal of Computational and Graphical Statistics 23 (2014): 1061-1079.

McElroy, T.S. and Holan, S.H. (2014) "Asymptotic theory of cepstral random fields." Annals of Statistics. 42: 64-86.

Moehl, J., Masters Thesis. Comparing models of Demographic Subpopulations. University of Tennessee, 2014.

Nagle, N., B. Buttenfield, S. Leyk, and S. E. Spielman. "Dasymetric modeling and uncertainty." The Annals of the Association of American Geographers 104 (2014): 80-95.

Continued on page 7

Publications (Continued)

Porter, A., Holan, S.H., Wikle, C.K., and Cressie, N. (2014). "Spatial Fay-Herriot Models for small area estimation with functional covariates." Spatial Statistics, 10, 27-42.

Porter, A.T. and Oleson, J. (2014) "A CAR model for multiple outcomes on mismatched lattices." Spatial and Spatio-Temporal Epidemiology. 11: 79-88.

Porter, A.T., Wikle, C.K., and Holan, S.H. (2015) (To Appear Australian & New Zealand Journal of Statistics) "Small area estimation via multivariate Fay-Herriot models with latent spatial dependence."

Spielman, S. E. "Spatial collective intelligence? Accuracy, credibility and volunteered geographic information." Cartography and Geographic Information Science 41 (2014): 115-124.

Spielman, S. E., D. Folch, and N. Nagle. "Causes and patterns of uncertainty in the American Community Survey." Applied Geography 46 (2014): 147-157.

Spielman, S. E., and P. Harrison. "The co-evolution of residential segregation and the built environment at the turn of the 20th century: a Schelling model." Transactions in GIS 18 (2014): 25-45.

Spielman, S. E., and J. Logan. "Using high resolution population data to identify neighborhoods and determine their boundaries." Annals of the Association of American Geographers 103 (2013): 67-84.

Wasi, Nada, and Aaron Flaaen (forthcoming), "Record Linkage using STATA: Pre-processing, linking and reviewing utilities" The Stata Journal.

Wikle, C.K. (2015) "Modern perspectives on ptatistics for spatio-temporal data." WIRES Computational Statistics. 7:86-98.

Wu, G., Holan, S.H., Nilon, C.H., and Wikle, C.K. (2015) (To Appear in Annals of Applied Statistics) "Bayesian binomial mixture models for estimating abundance in ecological monitoring studies."

Wu, G., Holan, S.H., and Wikle, C.K. (2013) "Hierarchical Bayesian spatio-temporal Conway-Maxwell poisson models with dynamic dispersion." Journal of Agricultural, Biological, and Environmental Statistics. 18: 335-356.

Yang, W.H., Wikle, C.K., Holan, S.H., Sudduth, K., and Meyers, D.B. (2015) "Bayesian analysis of spatially-dependent functional responses with spatially-dependent multi-dimensional functional predictors." Statistica Sinica. 25: 205-223.

Zhuang, L. and Cressie, N. (2015) "Bayesian hierarchical statistical SIRS models." Statistical Methods and Applications, forthcoming (doi: 10.1007/s10260-014-0280-9).

Presentations

The following are presentations that were made by people working at the eight research nodes involved with the NSF-Census Research Network (NCRN) grant. To look at more events on our calendar, or to subscribe to our calendar, click here. (http://www.ncrn.info/events/calendar/month)

Chris Wikle, Interaction-based parameterizations for nonlinear dynamic spatio-temporal models. Department of Statistics, The Ohio State University, Columbus, OH, October 1, 2014. http://www.ncrn.info/event/chris-wikle-presents-statistics-colloquium-ohio-state

Scott Holan, "Bayesian Lattice Filters for Time-Varying Autoregression and Time-Frequency Analysis," University of Maryland, Baltimore County, October 2, 2014. http://www.ncrn.info/event/scott-holan-presents-university-maryland

Chris Wikle, Interaction-based parameterizations for nonlinear dynamic spatio-temporal models. Department of Statistics, Texas A&M, College Station, TX, October 10, 2014.http://www.ncrn.info/event/chris-wikle-presents-statistics-colloquium-texas-am

(Continued on page 8)

Presentations (Continued)

Chris Wikle, Feature space reduction via dimension expansion of "big data"time-series covariates. Department of Statistics, The Pennsylvania State University, State College, PA, October 22, 2014. http://www.ncrn.info/event/chris-wikle-presents-pennsylvania-state-university

Noel Cressie, "Statistical modelling of big, spatial, non-Gaussian data" presented the 2014 Knibbs Lecture, Statistical Society of Australia Inc., Canberra, Australia, 4 November 2014; http://www.ncrn.info/event/noel-cressie-presents-2014-knibbs-lecture

Scott Holan, Bayesian Lattice Filters for Time-Varying Autoregression and Time-Frequency Analysis, Colorado School of Mines, Golden CO, November 2014. http://www.ncrn.info/event/scott-holan-presents-applied-mathematics-and-statistics-colloquium-colorado-school-mines

Chris Wikle, Feature space reduction via dimension expansion of "big data" time-series covariates. Quantitative Psychology Department, University of Missouri, Columbia, November 7, 2014. http://www.ncrn.info/event/chris-wikle-presents-quantitative-colloquium-university-missouri

Noel Cressie, Invited seminar speaker, Université de Paris-Ouest, Nanterre, France, 4 December 2014; "Predictive inference for big, spatial, non-Gaussian data" http://www.ncrn.info/event/noel-cressie-presents-universite%C2%B4-de-paris-ouest-nanterre-france

Isaac Sorkin, University of Michigan, presented "Ranking Firms Using Revealed Preference" at the Michael Beauregard Seminar in Macroeconomics at the University of Michigan, and Changkeun Lee, University of Michigan, presented "Was the Great Depression Cleansing? Evidence from the American Automotive Industry, 1929-1935" at the Center for Economic Studies at the U.S. Census Bureau, December 10, 2014. http://www.ncrn.info/event/isaac-sorkin-presents-michael-beauregard-seminar-macroeconomics

Changkeun Lee, University of Michigan, presents "Was the Great Depression Cleansing? Evidence from the American Automotive Industry, 1929-1935" at the Center for Economic Studies Seminar at the U.S. Census Bureau, December 11, 2014. http://www.ncrn.info/event/changkeun-lee-presents-center-economic-studies-us-census-bureau

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