FCSM Computational Statistics and the Production of Official Statistics Webinar on Blended Data

Friday, May 1 3:00-4:30pm (Eastern Time)

Registration is free, but required. We will email Zoom log-in information to registered users prior to the event.

Register at https://www.eventbrite.com/e/fcsm-cspos-webinar-on-blended-data-tickets-103262105738

While the larger April 2020 FCSM Research and Policy Conference has been cancelled, the FCSM interest group on Computational Statistics and the Production of Official Statistics (CSPOS) invites you to tune in to its scheduled May 1st meeting at 3:00pm. We'll be featuring the four contributed presentations on blended data that authors from CSPOS had prepared for the April conference. Agenda and speaker biographies and abstracts below.

Agenda

- *3:00—Welcome by moderator, Nathan Cruze, USDA NASS*
- *3:05—**Public Libraries and Collective Efficacy: An Exploratory Study of Blending Data from the Public Libraries Survey and the American Housing Survey**, Lisa Frehill, Institute of Museum and Library Services*
- *3:25—**Blending Privately-Provided Payroll Data and Government Statistics**, Leland Crane, Federal Reserve Board of Governors*
- *3:45—**Linking Official Statistics and Remote Sensing Data for Training Crop Yield Regression Models**, Luca Sartore, NISS and USDA NASS*
- *4:05—**Measuring the Cost and Impact of Open Source Software innovation on GitHub**, José Bayoán Santiago Calderón and Brandon Kramer, University of Virginia Biocomplexity Institute and Initiative*
- *4:30—Webinar concludes*

Presenter Biographies

Lisa M. Frehill has over 25 years of experience in academic, non-profit, and government settings. She earned her doctoral degree in Sociology and Systems and Industrial Engineering from the University of Arizona. As the IMLS Senior Statistician, Dr. Frehill publishes datasets and data file documentation, along with technical reports and research about museums, libraries, and other cultural organizations.

Leland Crane is a Senior Economist in the Industrial Output section of the Federal Reserve Board of Governors. His interests include search & matching models of the labor market, machine learning, and economic measurement using non-traditional data.

Luca Sartore is a NISS postdoctoral fellow, working in the Research and Development Division at USDA's National Agricultural Statistical Service. He received his Ph.D. in statistical sciences from the University of Padua. He was a former research fellow at the European Centre for Living Technology. Dr. Sartore's research focuses primarily on non-standard regression techniques and spatio-temporal models.

José Bayoán Santiago Calderón is a postdoctoral research associate at the Biocomplexity Institute and Initiative at the University of Virginia in the Social & Decision Analytics Division. He received his doctorate in economics from Claremont Graduate University in 2019. His specializations include econometrics and computational economics. Bayoán has worked as a research consultant, data scientist, and software developer in multiple fields. He is an active member of the Julia language community and regular contributor to open-source software.

Brandon L. Kramer is a postdoctoral research associate at the Biocomplexity Institute and Initiative at the University of Virginia in the Social & Decision Analytics Division. He recently completed his doctoral training in the Department of Sociology at Rutgers University. Dr. Kramer's research employs computational methods to examine topics at the intersection of science, technology, and health. In addition to his work on open-source software, Brandon is interested in how biomedical scientists conceptualize race, gender and other forms of diversity in their research.

Public Libraries and Collective Efficacy: An Exploratory Study of Blending Data from the Public Libraries Survey and the American Housing Survey

Presenter: Lisa Frehill, Institute of Museum and Library Services

There has been much research about the roles played by public libraries in enhancing social capital and civic engagement (Audunson et al 2007, Gong et al 2008, Johnson 2010 and 2012, Horrigan 2016, Vårheim 2014 and 2016). Most of these studies use analyses of relatively small numbers of individuals or libraries that provide often rich details about these cases. Our study explores using data from two separate Federal data collections, the Public Libraries Survey and the 2013 data in the Collective Efficacy module of the American Housing Survey to explore the extent to which these data sources can be blended to take a more macro view of the connection between libraries, social capital and civic engagement. Such an approach trades rich detail for broader generalizability and, therefore, potentially greater utility for more data users in the library data community. The paper reviews the methodological challenges associated with blending data from these disparate sources and provides lessons learned on how to overcome these challenges.

Blending Privately-Provided Payroll Data and Government Statistics

Presenter: Leland Crane, Federal Reserve Board of Governors

We combine employment information from three sources in order to accurately measure the U.S. labor market in real time. The data sources are (1) microdata from the payroll processing firm ADP, (2) the CES survey from BLS, and (3) administrative QCEW data, also from the BLS. While the ADP data cover a very large fraction of U.S. employment (the sample is similar in size to the CES), they are a convenience sample, not a probability sample. We use QCEW data to reweight the ADP data based on a set of observables to attenuate selection bias. We argue that pooling the CES and ADP data can reduce the sampling error inherent in both data sources. In particular, we infer "true" unobserved payroll employment growth using a state-space model that treats both series as noisy signals, and evaluate the forecasting power of various measures derived from this setup. We also incorporate the QCEW by benchmarking our ADP data to the QCEW at low frequencies, taking advantage of its comprehensive coverage of employment.

Linking Official Statistics and Remote Sensing Data for Training Crop Yield Regression Models

Presenter: Luca Sartore, National Institute of Statistical Sciences and USDA National Agricultural Statistics Service

USDA's National Agricultural Statistics Service (NASS) publishes more than 500 reports every year. Such publications include monthly and annual yield forecasts and estimates for major crops. To produce the forecasts, several surveys are conducted during the growing season. To enhance the accuracy of these forecasts, NASS uses external sources of information such as weather and remote sensing data obtained at different temporal and spatial resolutions by other national agencies. In fact, weather stations and satellites have provided additional information for the estimation of overall crop productivity, i.e. yield and planted area. The measurements from the Moderate Resolution Imaging Spectroradiometer (MODIS) project provide multispectral composite data at 250 m spatial resolution on crop regions throughout the growing season. An aggregation method based on empirical densities is proposed to reduce the loss of information. These densities are then linked to historical NASS official statistics with the aim to train regression models.

Measuring the Cost and Impact of Open Source Software Innovation on GitHub

Presenters: José Bayoán Santiago Calderón and Brandon L. Kramer, University of Virginia Biocomplexity Institute and Initiative Social and Decision Analytics Division

Open Source Software (OSS), defined by Open Source Initiative, is computer software with its source code shared with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose. OSS is developed, maintained, and extended both within and outside of the private sector, through the contribution of independent developers as well as people from universities, government research institutions, businesses, and nonprofits. Examples include Apache server software, and R statistical programming software. Despite its ubiquity and extensive use, reliable measures of the scope and impact of OSS developed outside of the business sector are scarce. Activities around OSS development, a vital component of science activity, are not well-measured in existing federal statistics on innovation. Many of the OSS projects are developed and maintained in free repositories, such as GitHub, and information embedded in these repositories, including the code, contributors, and development activity, is publicly available. In this paper, we use data from GitHub, the largest platform with 31 million users and developers worldwide, obtaining information about OSS projects. We collect 5.2 million project repositories, containing metadata such as author, license, commits (approved code edits), and lines of code. We adopt methods used in software engineering to estimate the resource cost associated with creating OSS. We use lines of code as the measure of effort to estimate the time spent on software development and calculate the monetary value using the average compensation for computer programmers from Bureau of Labor Statistics wage data and other costs based on national accounts methodologies. Finally, use network analysis methods developed for bibliometrics and patent analysis to study the impact of these projects.