REPRODUCIBILITY IN THE PRESENCE OF CONFIDENTIAL DATA

April 12, 2017

Abstract

Reproducibility is important. Confidential data - data that cannot be publicly distributed in the form that the original author used it - is often the topic of criticism when used in economic research, as it is perceived to not lend itself to reproducible research. In this session, we outline why that need not be the case, and why, in fact, restricted-access environments can in fact be at the vanguard of reproducible research.

Primary JEL Classification: A1 - General Economics

Secondary JEL Classification: B4 - Economic Methodology

If accepted, I would like the session considered for the AER Papers and Proceedings issue in May 2017, and understand that if selected for the P & P, this would impose significant page restrictions on all final papers: ***Yes***

Vilhuber acknowledges direct support from NSF Grant SES-1131848 (NCRN) and a grant from the Alfred P. Sloan Foundation.

Contents

1 Paper 1 **3**
- 1.1 Assessing Effective Reproducibility at the American Economic Journal: Applied Economics . . . 3
- 1.2 Discussant: Ian M. Schmutte . . . 3

2 Paper 2 **4**
- 2.1 Implementing Support for Reproducibility in a Restricted Access Environment: Case of the Federal Statistical Research Data Centers . . . 4
- 2.2 Discussant: Victoria Stodden (To be confirmed) . . . 4

3 Paper 3 **5**
- 3.1 How Amazon and Tripadvisor can help central banks . . . 5
- 3.2 Discussant: TBD . . . 5

Session Organizer: Lars Vilhuber, Cornell University

Email (lars.vilhuber@cornell.edu)

Chair: Jörg Heining, Institute for Employment Research (IAB, Germany)

Email (joerg.heining@iab.de)

1 Paper 1

1.1 Assessing Effective Reproducibility at the American Economic Journal: Applied Economics

Authors Lars Vilhuber (lars.vilhuber@cornell.edu) (Corresponding author, Cornell University), Flavio Stanchi (fs379@cornell.edu) (Cornell University), Hautahi Kingi (hrk55@cornell.edu) (IMPAQ International), Sylverie Herbert (sh2258@cornell.edu) (Cornell University)

One sentence description Success and failure of reproducibility in a journal

Abstract We describe effective reproducibility - the ability to actually download and re-execute the archives as provided by the authors - for a single journal with a replication policy, and analyze the causes for non-reproducibility. Reproducibility fails primarily when data are not accessible, but also fails because authors are not able to convey what protocols are available to reproduce their work. Furthermore, in all cases, authors failed to provide programs that link back to the original data.

Keywords Reproducibility; Confidential Data.

1.2 Discussant: Ian M. Schmutte

Email (schmutte@uga.edu)

2 Paper 2

2.1 Implementing Support for Reproducibility in a Restricted Access Environment: Case of the Federal Statistical Research Data Centers

Authors Lucia S Foster (lucia.s.foster@census.gov) (corresponding author), Shawn D Klimek (shawn.d.klimek@census.gov), Danielle Sandler (danielle.h.sandler@census.gov), Lars Vilhuber (lars.vilhuber@cornell.edu) (Cornell University)

One sentence description This paper describes an effort at the U.S. Census Bureau to support reproducibility of research performed using confidential Census microdata.

Abstract Replication of existing empirical work is one of the cornerstones of robust scientific inquiry. This paper describes an effort at the U.S. Census Bureau to support reproducibility of research performed using confidential Census microdata. This effort has three major components: documentation, access, and dissemination. The output from projects using Census microdata is already controlled and documented for disclosure avoidance reasons. We will leverage this existing structure to provide better tools for researchers using Census data to document the data and code they use for their research papers. As part of this additional documentation, we will clarify and improve the process by which their data and code can be accessed for replication purposes. The Federal Statistical Research Data Center (FSRDC) system provides access to Census microdata to an increasing number of researchers across the country. We plan to provide access to researchers who wish to conduct replications through a shorter and simpler process than the typical FSRDC research project proposal. We plan to disseminate the results of the replication studies, regardless of the outcome of those studies, as an ongoing CES Replication Studies series. This will add transparency to the process, increase the exposure of these studies, and reduce the effects publication bias. The goal of this project is to improve the understanding of the U.S. economy by making the research conducted by Census internal researchers and in the FSRDCs robust and replicable by the research community.

Keywords Reproducibility; Confidential Data.

2.2 Discussant: Victoria Stodden (To be confirmed)

3 Paper 3

3.1 How Amazon and Tripadvisor can help central banks

Authors Stefan Bender (stefan.bender@bundesbank.de) ((corresponding author, Deutsche Bundesbank) and Julia Lane (julia.lane@nyu.edu) (New York University)

One sentence description This paper describes how the application of metadata documentation techniques from Amazon and Trip Advisor have improved search, discovery and reproducibility of confidential microdata.

Abstract The inability to search and discover data and code is a major challenge for reproducible research. Similar challenges have been addressed in the private sector, particular by companies like Amazon and Tripadvisor. This paper describes the development of new approaches to metadata documentation and curation which draw on the techniques used by such companies and apply them to financial microdata held by central banks. The case is particularly relevant since banks collect very detailed microdata about for example - banks, non-financial institutes and/or securities, which are needed to analyze causal relationships. Although there is often not a great deal of metadata documentation, researcher access is critical to enable high quality analysis. We describe the development of new software platforms, inspired by Amazon and Tripadvisor, that engages all users in the systematic documentation of all relevant steps in a data generating process.

Keywords reproducibility, curation, confidential data, financial data, big data, data access, software solution.

3.2 Discussant: TBD