Replication Archive for "Reassessing the Spatial Mismatch Hypothesis," by David Card, Jesse Rothstein, and Moises Yi

Overview

This replication package contains the code to generate the analysis file and results for the above paper. It is divided into two parts:

- "internal" contains code that runs on the Census Bureau's secure data system.[1] This has two sub-components:
    - "build" prepares the main analysis files. This takes several days to run all the way through.
    - "analysis" conducts the analysis on the resulting files. Parts of this are also slow, but less so than the build package.
- "external" constructs the tables and figures in the paper from results generated by the analysis programs and disclosed through the Census review process.

The "internal" code relies on confidential data that can be accessed only inside the Census Bureau and is not available to outside researchers (even those with Census Special Sworn Status). Results from these analyses were transferred by hand into spreadsheets that were then disclosed through the Census review process. These spreadsheets are included in the "external/disclosed" folder[2]; the code in the "external/code" folder refers only to results in these sheets.

The code contained here also produces all of the results in the online appendix that are generated from internal Census data. For brevity, this package does not contain code to produce the appendix tables (A-2, A-3, A-6, and part of A-4) that are based solely on public-use ACS data.

Data availability

The paper relies exclusively on confidential microdata from the U.S. Census Bureau (Bureau of the Census 2018a,b). The files are only available to Census internal researchers, and were accessed under DMS Project 6000266 (subproject 7515812). Although variants of the LEHD data are available to external researchers with Census Special Sworn Status working on approved projects, these are different versions than the ones used for this project. The LEHD files used for this project are available only to internal Census Bureau researchers, and cannot be obtained by external researchers, even those with Special Sworn Status.

[1] The Census Bureau has ensured appropriate access and use of confidential data and has reviewed this code for disclosure avoidance protection (project 6000266: CBDRB-FY24-CES024-003).

[2] The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (project 6000266: CBDRB-FY24-CES024-001/002).

The archive includes the programs that we used to analyze the confidential Census data and the results of those programs that were cleared for release by the Census disclosure review process. The programs drew on the following files, available only inside the Census secure computing environment:

- jhf_interleave.sas7bdat
- ecf_seinunit_interleave.sas7bdat
- icf_us.sas7bdat
- vpers2001_1yr.sas7bdat
- vpers2002_1yr.sas7bdat
- vpers2003_1yr.sas7bdat
- vpers2004_1yr.sas7bdat
- vpers2005_1yr.sas7bdat
- vpers2006_1yr.sas7bdat
- vpers2007_1yr.sas7bdat
- vpers2008_1yr.sas7bdat
- vpers2009_1yr.sas7bdat
- vpers2010_1yr.sas7bdat
- vpers2011_1yr.sas7bdat
- vpers2012_1yr.sas7bdat
- vpers2013_1yr.sas7bdat
- vpers2014_1yr.sas7bdat
- vpers2015_1yr.sas7bdat
- acs2016_vpers_1yr.sas7bdat
- acs2017_vpers_1yr.sas7bdat
- crosswalk_acs2001.sas7bdat
- crosswalk_acs2002.sas7bdat
- crosswalk_acs2003.sas7bdat
- crosswalk_acs2004.sas7bdat
- crosswalk_acs2005.sas7bdat
- crosswalk_acs2006.sas7bdat
- crosswalk_acs2007.sas7bdat
- crosswalk_acs2008.sas7bdat
- crosswalk_acs2009.sas7bdat
- crosswalk_acs2010.sas7bdat
- crosswalk_acs2011.sas7bdat
- crosswalk_acs2012.sas7bdat
- crosswalk_acs2013.sas7bdat
- crosswalk_acs2014.sas7bdat
- crosswalk_acs2015.sas7bdat
- crosswalk_acs2016.sas7bdat
- crosswalk_acs2017.sas7bdat

The first three files contain LEHD data, as well as geographic information on LEHD observations. We used versions of these files that contained data through 2018. The Census Bureau updates these files annually with new records; our code should work with the updated data, as it excludes post-2018 observations. The remaining files contain ACS data and information needed to link the ACS respondents to the LEHD.

Computational requirements

The repository includes scripts that govern all of the analysis. They should be run in the following order:

- internal/00_run_all.do. This must be run on the internal Census secure servers.
- external/publicmaster.do. This must have access to the results (generated by the above programs) that have been transferred by hand into spreadsheets, slightly redacted (e.g., by rounding), and disclosed for public release by Census. Note that the external/code folder also contains three spreadsheets (table1.xlsx, tableA1.xlsx, and tableA5.xlsx), that build the respective tables directly from the disclosed results, via Excel formulas.

Confidential components

The code in the first portion of the project was executed on the Census Bureau's RDC system. It used:

- Stata-MP (version 16.0)
- SAS (version 9.4)
- Matlab (version R2020a), with add-on package
    - MatlabBGL (version 4.0), available from https://www.cs.purdue.edu/homes/dgleich/packages/matlab_bgl/ or https://github.com/dgleich/matlab-bgl.
- R (version 4.3.1), with add-on packages haven, data.table, tidyverse, gdata, logr.

It takes several days to run all the way through, assuming no interruption.

Public components

The code in the second portion was executed on a Macbook Pro, with an Apple M1 Pro chip and 32 GB of RAM, running OSX version 13.6.3. It used Stata-MP, version 16.1. This portion of the code takes just a few moments to run.

Mapping of internal code output to disclosure tables

The internal code structure and dependencies are elaborated in a separate document, internal/notes.txt.

Note several of the internal programs have been redacted for disclosure. To run these programs, directory locations would need to be set in:

- Internal/00_directories.do
- Internal/build/0a_mig5_clean1_bw.sas
- Internal/build/2_prepresidlocs.sas
- Internal/build/firmAKM_callable.m
- Internal/analysis/gridmult.R
- Internal/analysis/gridcorrs.R

The disclosure spreadsheets that are used for the tables and figures of the paper were assembled by hand from the output of the programs in portion 1 of the project, run inside the Census RDC, as part of the disclosure process. This table describes which portions of the code in that portion generate the numbers in each of the tables in those spreadsheets:

Disclosure tab	Program (within internal/code directory)	Used in
Yi_tabs_T13T26_1, tab 1	Tables/AKMsummary.do	Table1.xlsx, TableA1.xlsx
Yi_tabs_T13T26_1, tab 2	Gridmulttable.do	Fig1.do
Yi_tabs_T13T26_1, tab 3	Psi_commute.do	Fig2.do
Yi_tabs_T13T26_1, tab 4	Commutedistn.do	FigA4.do
Yi_tabs_T13T26_1, tab 5	Gridcorrs.R	<Not used>
Yi_tabs_T13T26_2, tab 1	Tables/AKMsummary.do	TableA1.xlsx
Yi_tabs_T13T26_2, tab 2	Commutedensity.do	FigA2.do
Yi_tabs_T13T26_2, tab 3	Multiestab_commutedist.do	FigA3.do*
Yi_tabs_T13T26_2, tab 4	Gridmulttable.do	FigA1.do
Yi_tabs_T13T26_2, tab 5	Tables/elasticities.do	TableA5.xlsx
Yi_tabs_T13T26_2, tab 6	Created by hand; no data.	<Referred to in text>

* The results in Yi_tax_T13T26_2, tab 3 differ slightly from those created by the internal code, as they were created by an earlier version of the code. As a result, were FigA3.do to be applied to results generated by the code contained herein, it would differ slightly from that in the online appendix. We have verified that the changes are minimal and do not affect any of the conclusions we draw from the figure. We have opted not to disclose the updated version of the data for this figure due to the minimal impact of the changes and the fact that it appears only in the online appendix.

References

Bureau of the Census. “American Community Survey, 2001-2017” [database]. United States Department of Commerce [publisher]. (2018a; accessed December 12, 2023).

Bureau of the Census. “Longitudinal Employer Household Dynamics, 2010-2018” [database]. United States Department of Commerce [publisher]. (2018b; accessed December 12, 2023).