Replication package for "Structural Change and Internal Labor Migration: Evidence from the Great Depression"

Christopher Boone Laurence Wilse-Samson

August 28, 2021

1 Data Availability

Code for data cleaning and analysis is provided as part of this replication package. It is available at the Review of Economics and Statistics Dataverse.

This replication package contains the programs (Stata do-files) and data sets necessary to reproduce the tables and figures contained in the paper. We also provide programs for constructing these analysis datasets. Some of the raw input data used to construct these data sets is provided in this replication package; for other raw input data, we provide information and guidance for how to obtain these data.

1.1 Historical county-level data available from ICPSR

Much of the county-level data comes from two studies available at the ICPSR data archive:

1. ICPSR 2896: Historical, Demographic, Economic, and Social Data: The United States, 1790-2002 (Haines and ICPSR 2010)
2. ICPSR 35206: United States Agriculture Data, 1840 - 2012 (Haines, Fishback and Rhode 2018).

These files were downloaded from the website of the Inter-university Consortium for Political and Social Research (ICPSR) at https://www.icpsr.umich.edu/web/ICPSR/studies/2896/versions/V3 and https://www.icpsr.umich.edu/web/ICPSR/studies/35206/versions/V4. The relevant files are included in this replication package (in the folders `dataRAW/ICPSR_02896/` and `dataRAW/ICPSR_35206/`).

1.2 Complete Count Population Census Data (1930, 1940)

We make use of the complete count decennial population census data for 1930 and 1940. These data have restricted access. These data were originally digitized by Ancestry.com and have been made available to researchers as a result of a collaboration between Ancestry.com and IPUMS-USA (Ruggles et al. 2020). For the results in this paper, we used the data sets hosted by NBER, which can be made available to NBER affiliates. (See `create_individual_level_data.do` for information on the specific versions of the data sets used.) Other researchers interested in accessing the restricted access data can contact

IPUMS for information on how to obtain access; see the section on "Restricted Versions" at the following url: https://usa.ipums.org/usa/complete_count.shtml. We wish to acknowledge Ancestry.com for providing the underlying data making this research possible.

Anonymized versions of the complete count censuses are hosted by IPUMS-USA (see https://usa.ipums.org/usa/complete_count.shtml); those data do not includes names, so they cannot be used to link individuals by name, but they can be merged with data from the Census Linking Project (Abramitzky, Boustan and Rashid 2020) in order to do so. (See the subsection on the Census Linking Project below for more information.)

1.3 Data on county-level slope or gradient of the land

Data on the slope of the land was obtained from the Digital General Soil Map of the United States or STATSGO2 Database (USDA Soil Survey Staff 2016). The file wss_gsmsoil_US_[2016-10-13].zip was downloaded in December 2018 from USDA Web Soil Survey at http://websoilsurvey.nrcs.usda.gov/. To obtain the average slope or gradient by county, the following procedure was used: first open the database file using Microsoft Access, link to tables, and save; then open shp file using ArcMap; in ArcMap, add the 1930 county boundaries using the NHGIS county boundary file (see below); perform a spatial join (on mukey, using the "component" table, which contains the slope fields). The average gradient by county is exported to the text file `Export_Output_statsgo2_1930counties.txt`, which is included in this replication archive (in the `dataRAW/` directory).

1.4 Data from Fishback, Horrace and Kantor (2005) and Fishback et al. (2011)

We obtain the data set provided by Fishback et al. (2011). We use data on weather, along with data on employment and output in the retail, wholesale, and manufacturing sectors. The original data sources are described in detail in the appendix to Fishback et al. (2011). We originally obtained the data from https://econ.arizona.edu/weather-demography-economy-and-new-deal-county-level-1930-1940, but that link is no longer active. The same data set is now hosted at https://www.openicpsr.org/openicpsr/project/101199/version/V1/view (under the Weather–Demography–and-the-New-Deal directory). We also include the relevant variables in this replication package.

We also use data on county-level per capita AAA spending from Fishback, Horrace and Kantor (2005). These data are available at https://www.openicpsr.org/openicpsr/project/101199/version/V1/view (under the Retail-Sales directory).

The relevant files are included in this replication archive in the following directory: `dataRAW/fishback/`.

1.5 County Border Fixes and Erosion Data

We use the data files produced by Hornbeck (2010) to account for changes in county borders over time. These correspondences can be obtained from https://voices.uchicago.edu/richardhornbeck/ (see link to BWreplication_BorderFixes.rar under "Replication Files and Border Fixes"). The relevant files are included in this replication archive in the following directory: `dataRAW/border_changes/`.

We also use county-level erosion information from Hornbeck (2012). These data are available to download at https://voices.uchicago.edu/richardhornbeck/. The relevant data set is provided here in `dataRAW/erosion_1910.txt`.

1.6 Historical Statistics of the United States

We obtain several data series from the *Historical Statistics of the United States*, in particular the series produced by Ferrie (2006), Olmstead and Rhode (2006*a*), Olmstead and Rhode (2006*b*), and Sutch (2006). The relevant data are included in this replication archive (`dataRAW/hsus/`).

1.7 Commodity Price Data from Blattman, Hwang and Williamson (2007)

We use data on annual commodity prices from Blattman, Hwang and Williamson (2007). These data can be downloaded from https://chrisblattman.com/research/ (see link to "Commodity price indices 1865-1950.zip" under "full source data on commodity prices 1865-1950"). The file `Commodity price indices 1865-1950.xls` is included in this replication archive (in `dataRAW/`).

1.8 Census Samples from IPUMS-USA

We obtain data from the population censuses of 1900 (1%), 1910 (1%), 1920 (1%), 1930 (5%), and 1940 (1%) from IPUMS-USA (Ruggles et al. 2020).

The 1900, 1910, and 1920 samples are used to calculate the county-level farm population in those years. The corresponding county-level statistics are provided in `ipums_population_farm_1900.dta`, `ipums_population_farm_1910.dta`, and `ipums_population_farm_1920.dta` (in `dataRAW/`).

The 1930 and 1940 samples are used to construct the county-level (Bartik) predicted change in employment. These statistics are provided in `dataRAW/bartik_ipums_county_1930_1940.dta`.

The 1930 sample is used to construct county-level measures of the percent of manufacturing workers in durables-producing industries, as well as the percent of workers in the manufacturing sector (`dataRAW/ipums_durable_county_1930.dta`).

1.9 Crop Suitability

We obtain measures of crop suitability for 8 crops (cotton, maize, oats, sugar beet, sugar cane, tobacco, wet rice, and wheat) from the Global Agro-ecological Zones project at http://www.gaez.iiasa.ac.at/. We use the crop suitability index (value) for rainfed, intermediate inputs, baseline time period, and join the GAEZ data with county boundaries from NHGIS. The resulting county-level data are included in this replication package in `dataRAW/crop_suitability.dta`.

1.10 County Boundary Files from NHGIS

We download shpfiles containing historical U.S. county boundaries from the National Historical Geographic Information System (NHGIS) Database (Manson et al. 2018). These boundaries are used to compute the average slope by county discussed above. We also use them to

compute neighboring counties (see dataRAW/CountyNeighbors1930_Export_Output.txt). Stata data sets containing the boundary information are included in this replication archive (dataRAW/ctycoord1930.dta and dataRAW/ctydb1930.dta) and used to produce the maps in the paper.

1.11 NASS Farm Employment

We use data from the series on Average Number of Workers on Farms, United States, 1910-1974, from the National Agricultural Statistics Services. Specifically, we use the information on family workers and hired workers. The raw data are provided in this replication archive in the dataRAW/nass/ directory. The files flbulwks.wk1, frmlabor.txt, and README.DOC were downloaded in May 2014 from the USDA archive hosted by Cornell University Library; the url is no longer active, so we include the raw files here. We opened the file flbulwks.wk1 and exported it to excel format as flbulwks.xlsx (also provided); flbulwks_farmworkerswages.csv contains a cleaned version of these data for importing into Stata.

1.12 Census Linking Project

To compare our main results to those using alternative linking procedures, we make use of the crosswalks provided by the Census Linking Project (Abramitzky, Boustan and Rashid 2020). We downloaded the 1930-1940 crosswalks from https://censuslinkingproject.org/data/ and combined these with the 100% 1930 and 1940 census data from IPUMS. See codeSTATA/create_match_CLP.do for more information, including the code for creating the data sets for our analysis and the tables for the paper.

2 Dataset list

2.1 dataSTATA/

These datasets are included in the dataSTATA/ folder. See master_get_data.do for the do-files used to construct these data sets.

1. individual_level_data.dta. This is the main individual-level data set used for the analysis. Observations are individuals who have been linked between the 1930 and 1940 population censuses. The code to create this data set is provided in create_individual_level_data.do; however, to run that code, the user first needs to acquire access to the complete count 1930 and 1940 censuses.
2. county_level_data.dta. This is the main county-level data set used in the analysis. The user can execute create_county_level_data.do to recreate this data set.
3. long_panel.dta. A county-level data set used for some of the tables and figures. The user can execute create_long_panel.do to recreate this data set.

The user can recreate the following data sets by running clean_raw_datasets.do:

4. statsgo2_slope.dta. A county-level data set containing the slope measure used in the analysis.
5. erosion_map_1930.dta. A county-level data set containing information on erosion.

6. `mfg_data_wide.dta`. A county-level data set containing information on employment and other characteristics for the retail, wholesale, and manufacturing sectors.
7. `d_lnmfgave_1929_1933.dta`. A county-level data set containing the variable `d_lnmfgave_1929_1933` (the change in log manufacturing employment between 1929 and 1933).
8. `cropAcreShares1910.dta`. A county-level data set containing share of farmland in specific crops in 1910.
9. `fishback_aaa.dta`. A county-level data set containing AAA spending per capita.
10. `fishback_weather.dta`. A county-level data set containing various weather-related variables.
11. `prices_hsus_1928_1932_2.dta`. A county-level data set containing annual average crop price growth between 1928 and 1932 for specific crops (from HSUS data).
12. `price_shock_depression_2.dta`. A county-level data set containing the crop price shock measure (a weighted average of the price change measure from `prices_hsus_1928_1932_2.dta`, where the weights are based on the share of farmland devoted to each crop).
13. `commodity_price_changes_3.dta`. A data set containing the annual average growth rate of 7 crops from 1928 to 1932 (from Blattman et al. data).
14. `gdp_by_decade.dta`. A data set containing the average annual GDP growth by decade (from HSUS data); see `codeSTATA/clean_raw_datasets.do`.

2.2 dataRAW/

This directory contains input data that is necessary to construct our analysis datasets.

1. `bartik_ipums_county_1930_1940.dta`. This file contains county-level Bartik-predicted change in employment between 1930 and 1940, based on the 1930 5% and 1940 1% IPUMS census samples, and using the IPUMS industrial classifications.
2. `ipums_durable_county_1930.dta`. This file contains county-level measures of the percent of total employment in manufacturing, and the percent of manufacturing workers in durables-producing industries. Constructed using the 1930 5% census sample and the IPUMS classifications for durables vs. non-durables.
3. `Commodity price indices 1865-1950.xls`. Contains commodity price data from Blattman, Hwang and Williamson (2007). These data can be downloaded from https://chrisblattman.com/research/.
4. `corr_stateabbrevfips.dta`, `corr_statedivisionfips.dta`, `corr_stateicpsrfips.dta`, and `icpsr_fips.dta`. These files contain correspondences between (respectively) state FIPS codes and state abbreviations, state divisions and state FIPS, state ICPSR codes and state FIPS codes, and ICPSR state/county codes and county FIPS.
5. `CountyNeighbors1930_Export_Output.txt`. A file listing all neighboring county pairs using 1930 boundaries. Exported from ArcMap using the NHGIS 1930 shpfile.
6. `ctycoord1930.dta` and `ctydb1930.dta`. Output of the -shp2dta- Stata command using the 1930 NHGIS county shpfile.

7. `crop_suitability.dta`. Contains county-level crop suitability data, based on the suitability data obtained from GAEZ.

8. `erosion_1910.txt`. County-level erosion information from Hornbeck (2012). These data are available to download at https://voices.uchicago.edu/richardhornbeck/.

9. `Export_Output_statsgo2_1930counties.txt`. A file containing average slope by county. Obtained by joining the slope information in the STATSGO2 database to the 1930 NHGIS county shpfile.

10. `ipums_population_farm_1900.dta`, `ipums_population_farm_1910.dta`, `ipums_population_farm_1920.dta`. County-level estimates of the farm population in 1900, 1910, and 1920 from the IPUMS census samples for those years.

2.3 dataRAW/border_changes/

This directory contains the files used to adjust the data for changes in county borders over time, from Hornbeck (2010). These correspondences can be obtained from https://voices.uchicago.edu/richardhornbeck/.

2.4 dataRAW/fishback/

This directory contains the data we use from Fishback, Horrace and Kantor (2005) and Fishback et al. (2011), available for download from https://www.openicpsr.org/openicpsr/project/101199/version/V1/view.

1. `fishback_RetailsalesDataSet.csv`. These are the data on Retail-Sales from Fishback, Horrace and Kantor (2005). We make use of the information on AAA spending by county.

2. `mfgvars_fromclimatemortality.dta`. This file contains a subset of the Weather–Demography–and-the-New-Deal data from Fishback et al. (2011), including the information on county-level employment and output in the retail, wholesale, and manufacturing sectors.

3. `weather_fromclimatemortality.dta`. This file contains a subset of the Weather–Demography–and-the-New-Deal data from Fishback et al. (2011), including the information on county-level weather.

4. `county_names_and_codes.csv`. This file contains a correspondence between county codes and county names that we use for merging some of the above data.

2.5 dataRAW/hsus/

This directory contains several data series from the *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition.*

2.6 dataRAW/ICPSR_02896/

This directory contains the relevant files from ICPSR study no. 02896, the project on Historical, Demographic, Economic, and Social Data: The United States, 1790-2002 (Haines and ICPSR 2010).

2.7 dataRAW/ICPSR_35206/

This directory contains the relevant files from ICPSR study no. 35206, the project on United States Agriculture Data, 1840 - 2012 (Haines, Fishback and Rhode 2018).

2.8 dataRAW/nass/

This directory contains the historical data on farm employment from NASS. See above for details.

3 Computational requirements

3.1 Software requirements

- Stata (code was last run with Stata/MP version 16.1)
  - There are a number of user-defined commands that must be installed for the code to run. The programs `codeSTATA/master.do` and `codeSTATA/master_get_data.do` will attempt to install them (if they are not already installed) by calling the program `codeSTATA/install_user_written_programs.do`. The following commands are required:

    reghdfe (Correia 2016*b*, 2014), spmap (Pisati 2007), ftools (Correia 2016*a*), estout (Jann 2004), winsor (Cox 1998), missings (Cox 2015), binsreg (Cattaneo et al. 2021).
  - This code was last run using the latest version of these programs as of August 2021. Future updates to these programs may change their behavior.

3.2 Description of programs

- Programs in the `codeSTATA/` directory reproduce the tables and figures from the paper. The program `codeSTATA/master.do` prepares Stata, loads required programs, sets key globals, and calls the other do-files to create tables and figures.
- Programs are also provided to create the analysis datasets from raw input data. The program `codeSTATA/master_get_data.do` coordinates this process and calls the individual do-files. Some commands in `codeSTATA/master_get_data.do` are commented out because additional work is required prior to running them. Since the analysis data sets are already provided in this replication package, it is not necessary to run `codeSTATA/master_get_data.do` to reproduce the tables and figures; see `codeSTATA/master.do` instead.

3.3 Computation time

The programs as provided will execute relatively quickly. `master.do` will run in about 30 mins, and `master_get_data.do` will run in under 1 minute. Some of the commands that are commented out (because they will not run without additional data not provided here) take more time to execute: `create_individual_level_data.do` can take several hours, and the descriptive statistics table in `tables_appendix.do` can also take several hours.

4 Instructions

1. Download the files in the replication archive and reproduce the folder structure. It should look like the following:
   - codeSTATA/
   - dataRAW/
     - border_changes/
     - fishback/
     - hsus/
     - ICPSR_02896/
     - ICPSR_35206/
     - nass/
   - dataSTATA/
   - figures_tables/
   - temp/
2. Update the global macro BASEDIR in `codeSTATA/master.do`. Set this path to the top-level folder containing `codeSTATA/`, `dataRAW/`, `dataSTATA/`, etc.
3. Run `codeSTATA/master.do` to replicate the tables and figures. Once this is completed, you can compile `figures_and_tables.tex` in order to view the output. All of the datasets required to run `codeSTATA/master.do` are provided in this replication archive.
4. (optional) The program `codeSTATA/master_get_data.do` can be used to reproduce the intermediate datasets provided in this archive. This do-file calls the other do-files contained in the `codeSTATA/` directory. Most of those do-files will run immediately, except for the file `create_individual_level_data.do`, which is commented out and requires additional data not contained in the replication archive; see above for information on accesssing the additional data necessary.

5 List of tables & figures and corresponding programs

The program `codeSTATA/master.do` produces all of the tables and figures in the paper (except for part of Appendix Table A1, and Appendix Tables C1-C3, which require the user to first acquire additional data). The table below lists the individual do-files that create each table or figure. All of these do-files are called by `codeSTATA/master.do`, and the output is saved in the `figures_tables/` directory.

Table or Figure	**Program**
Figures 1-6	`codeSTATA/figures.do`
Appendix Figures A1-A10	`codeSTATA/figures_appendix.do`
Appendix Figures D1-D3	`codeSTATA/figures_appendix.do`
Tables 1-6	`codeSTATA/tables.do`

Table or Figure	**Program** (cont.)
Appendix Tables A1-A18	`codeSTATA/tables_appendix.do`
Appendix Tables B1-B2	`codeSTATA/tables_appendix.do`
Appendix Tables C1-C3	`codeSTATA/tables_appendix.do`
Appendix Table D1	`codeSTATA/tables_appendix.do`
Appendix Tables E1-E2	`codeSTATA/tables_appendix.do`

6 References

References

Abramitzky, Ran, Leah Boustan, and Myera Rashid. 2020. "Census Linking Project: Version 1.0 [dataset]." https://censuslinkingproject.org.

Blattman, Christopher, Jason Hwang, and Jeffrey G. Williamson. 2007. "Winners and losers in the commodity lottery: The impact of terms of trade growth and volatility in the Periphery 1870–1939." *Journal of Development Economics*, 82(1): 156 – 179.

Cattaneo, Matias D., Richard K. Crump, Max H. Farrell, and Yingjie Feng. 2021. "Binscatter Regressions." Working Paper.

Correia, Sergio. 2014. "REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects." *Statistical Software Components, Boston College Department of Economics.*

Correia, Sergio. 2016*a*. "FTOOLS: Stata module to provide alternatives to common Stata commands optimized for large datasets." *Statistical Software Components, Boston College Department of Economics.*

Correia, Sergio. 2016*b*. "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator." Working Paper.

Cox, Nicholas J. 1998. "WINSOR: Stata module to Winsorize a variable." *Statistical Software Components, Boston College Department of Economics.*

Cox, N. J. 2015. "Speaking Stata: A set of utilities for managing missing values." *Stata Journal*, 15(4): 1174–1185(12).

Ferrie, Joseph P. 2006. "Change in the farm population through births, deaths, and migration: 1920-1970. Table Ac414-418." In *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition.* , ed. Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch and Gavin Wright. New York:Cambridge University Press.

Fishback, Price V., Werner Troesken, Trevor Kollmann, Michael Haines, Paul W. Rhode, and Melissa Thomasson. 2011. "Information and the Impact of Climate and Weather on Mortality Rates During the Great Depression." In *The Economics of Climate Change: Adaptations Past and Present. NBER Chapters*, 131–167. National Bureau of Economic Research, Inc.

Fishback, Price V., William C. Horrace, and Shawn Kantor. 2005. "Did New Deal Grant Programs Stimulate Local Economies? A Study of Federal Grants and Retail Sales during the Great Depression." *The Journal of Economic History*, 65(1): 36–71.

Haines, Michael, Price Fishback, and Paul Rhode. 2018. *United States Agriculture Data, 1840-2012 [Computer file].* Ann Arbor, MI:Inter-university Consortium for Political and Social Research [distributor], 2018-08-20.

Haines, Michael R., and ICPSR. 2010. *Historical, Demographic, Economic, and Social Data: The United States, 1790-2002 [Computer file]. ICPSR02896-v3.* Ann Arbor, MI:Inter-university Consortium for Political and Social Research [distributor], 2010-05-21.

Hornbeck, Richard. 2010. "Barbed wire: Property rights and agricultural development." *The Quarterly Journal of Economics*, 125(2): 767–810.

Hornbeck, Richard. 2012. "The Enduring Impact of the American Dust Bowl: Short- and Long-Run Adjustments to Environmental Catastrophe." *American Economic Review*, 102(4): 1477–1507.

Jann, Ben. 2004. "ESTOUT: Stata module to make regression tables." *Statistical Software Components, Boston College Department of Economics.*

Manson, Steven, Jonathan Schroeder, David Van Riper, and Steven Ruggles. 2018. *IPUMS National Historical Geographic Information System: Version 13.0 [Database].* Minneapolis, MN:University of Minnesota.

Olmstead, Alan L., and Paul W. Rhode. 2006*a*. "Crops and Livestock." In *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition.* , ed. Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch and Gavin Wright. New York:Cambridge University Press.

Olmstead, Alan L., and Paul W. Rhode. 2006*b*. "Farms - number, population, land, and value of property: 1910–1999 [Annual]. Table Da1-13." In *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition.* , ed. Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch and Gavin Wright. New York:Cambridge University Press.

Pisati, Maurizio. 2007. "SPMAP: Stata module to visualize spatial data." *Statistical Software Components, Boston College Department of Economics.*

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2020. *IPUMS USA: Version 10.0 [dataset].* Minneapolis, MN:IPUMS.

Sutch, Richard. 2006. "Gross domestic product: 1790–2002 [Continuous annual series]. Table Ca9-19." In *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition.* , ed. Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch and Gavin Wright. New York:Cambridge University Press.

USDA Soil Survey Staff. 2016. "U.S. General Soil Map (STATSGO2) [dataset]." *Available online at https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053629. Accessed 29 May 2018.*