Replication Package for
"Non-Random Exposure to Exogenous Shocks"
by Kirill Borusyak and Peter Hull

Overview

This package reproduces the analysis of the employment impacts of high-speed railroads (HSR) in China. The data construction and analysis are conducted in Stata and Python, and all the raw data is provided. Most of code can be run on any computer. One part (as described below) is better implemented on a research computing cluster on which you can run 100 jobs, ideally in parallel. If run on the cluster, the expected runtime is 3 hours.

This readme also describes Stata packages that we provide for other applications in the future: (1) to compute spatially clustered standard errors in instrumental variable regressions and (2) to perform randomization inference in recentered IV regressions.

Data Availability and Provenance Statements

Statement about Rights

- ☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the LICENSE.txt file.

Summary of Availability

- ☑ All data **are** publicly available.
- ☐ Some data **cannot be made** publicly available.
- ☐ **No data can be made** publicly available.

Details on each Data Source

Data Name	Data Files	Location	Provided
Population by Prefecture	Population.xlsx	raw/	YES
Characteristics of HSR Lines	stations.xlsx	raw/	YES

Chinese City Statistical Yearbooks	various	raw/yearbooks	YES
City Centroids	CityCentroids.xls	raw/gis	YES
Coordinates of Capitals	ProvCapitals.xls, PrefCapitals.xls	raw/gis	YES
Processing Instructions	chinese_words_translated.csv, city_name_errors.xlsx, extract_excel_file_list.xlsx, file_index.csv	raw/input	YES

Note: Data citations are given in the detailed descriptions below.

Population by Prefecture

Population by Chinese prefecture in 2000 and 2010
Sources:

Sheet "Data": The main data were copied from https://www.citypopulation.de/en/china/admin/ (accessed on Nov 20, 2018)

Sheet "HubeiCorrection": For the Hubei province the data are from https://www.citypopulation.de/en/china/hubei/admin/ (accessed at the same time).

Sheet "Manual": contains spelling corrections and was done manually by the authors.

Citations: Brinkhoff (2018)

Data files: raw/Population.xlsx

Characteristics of HSR Lines

The list and characteristics of HSR lines; the ordered list of cities on each line. This includes both opened lines and lines that have been planned or under construction as of April 2019.

Sources: Produced manually. Our starting points are Map 1.2 of Lawrence et al. (2019), China Railway Yearbooks (China Railway Yearbook Editorial Board, 2001-2013), and the replication files of Lin (2017). We cross-check network links across these sources and use Internet resources such as Wikipedia and Baidu Baike to confirm and fill in missing information.

Data files: raw/stations.xlsx

Chinese City Statistical Yearbooks

Chinese City Statistical Yearbooks, 2000-2017. Subfolders ending in _prefabove are for the data by prefecture and higher aggregation levels, while _clevel is for county-level information.

Sources:

For 2000-2015 excluding 2009 and 2011: http://oversea.cnki.net.proxy.uchicago.edu/kns55/default.aspx (accessed on Jan 23, 2019 via a University of Chicago portal).

For 2009, 2011, 2016, and 2017: http://tongji.oversea.cnki.net/chn/navi/HomePage.aspx?id=N2018050234&name=YZGCA&floor=1 (same as above)

As of 2023, these links are no longer operational. However, PDF versions of the Yearbooks can be accessed here: https://oversea.cnki.net/KNavi/YearbookDetail?pcode=CYFD&pykm=YZGCA&bh=&uniplatform=OVERSEA&language=en (and can be navigated to from the homepage https://oversea.cnki.net by searching 中国城市统计年鉴 and clicking Chinese City Statistical Yearbooks under the Yearbooks tab). If you have library access through an institution, try reloading this address with your library proxy. Otherwise, the PDFs can be purchased for $2 a page (each of the 18 yearbooks averages about 300 pages).

Citation: China Statistics Press (2000-2017)

Data files: Various Excel files in raw/yearbooks

City Centroids

List and centroid coordinates of prefectures

Sources: The data for the underlying the shapefiles was downloaded from https://data.humdata.org/dataset/china-administrative-boundaries (accessed on Apr 4, 2020). It was exported into Excel by geocoding in ArcGIS. The shapefiles are available in gis/rawmap/chn_adm2.

Citation: OCHA Regional Office for Asia and the Pacific (2020)

Data files: raw/gis/CityCentroids.xls

Coordinates of Capitals
Geographic coordinates of province and prefecture capitals

Sources: The data for the underlying shapefiles was downloaded from https://data.humdata.org/dataset/province-and-prefecture-capitals-of-china (accessed on Apr 4, 2020). It was exported into Excel by geocoding in ArcGIS. The shapefiles are available in gis/rawmap/capitals.

Citation: OCHA Regional Office for Asia and the Pacific (2018)

Data files: raw/gis/ProvCapitals.xls, raw/gis/PrefCapitals.xls

Processing Instructions Data
Instructions for processing Chinese City Statistical Yearbooks: which files and variables to read, etc.

Sources: Produced manually by the authors

Data files: chinese_words_translated.csv, city_name_errors.xlsx, extract_excel_file_list.xlsx, file_index.csv

Computational requirements

Software Requirements

- Stata (code was last run with version 17)
    - geodist (version distributed on 20190624)
    - `cleanchars` (version distributed on 20131013)
    - renvars (installed from Stata Journal volume 5 number 4)
    - packages are installed in part 1 of master_hsr.do, described below
- Python (code was last run with Python 3.9.12)
    - `pandas>=2.0.2`
    - `numpy`
    - `tqdm`
    - `xlrd`
    - `openpyxl`

    - `msoffcrypto-tool`
    - the file "`requirements.txt`" lists these dependencies, and they are installed automatically in part 4 of master_hsr.do
- ArcMap 10.8.2

Portions of the code use bash scripting, which may require Linux.

Controlled Randomness

Random seed is set on line 2 of code/build_data/5_reshuffle_line.do.

Memory and Runtime Requirements

The code was last run on a 4-core Intel-based laptop with MacOS version 12.6.7. The code run locally (not on the computing cluster) takes about 15 minutes to run.

Portions of the code were last run on a SLURM-based research computing cluster with Stata-MP installed as a module. 100 parallel jobs are run that take ~45 min each with 2 GB of memory. In the last run, it took 3 hours from the time the first job started to the last job finishing because of the availability of nodes. If you do not have access to a cluster, the jobs could be run locally on your computer (see details below).

Description of programs/code

Analysis Files:

The file master_hsr.do will run the complete analysis, and is broken up into 6 parts:

1. Set directories and install necessary packages
2. Prepare the railway database
    - 1_clean_population.do imports the population data
    - 2_clean_cities.do cleans the database of prefecture-level cities
    - 3_clean_lines.do cleans the database of lines and stations
    - 4_ma2007.do computes market access (MA) for the initial year 2007
    - 5_reshuffle_lines.do generates reshuffled lines
3. Run the server processing (code in *server/* folder)
    - run_server_2016_slurm.sh or run_server_2016_torque.sh (depending on your cluster environment as described below) calls 100 instances of process_scenarios_2016.do, which produces market access by city in 2016 for twenty simulated scenarios each time (so that in total there are

2,000 scenarios, including the one with the actual railroads by 2016). If you do not have access to the cluster, this step can be run on your local machine with appropriate patience; see details below.

4. Clean Chinese City Statistical Yearbooks in python
    - 6_unlock_excel_files.py unlocks the locked raw yearbook Excel files
    - 7_extract_excel.py reads the yearbooks into simple tables
    - 8_translate_file_index.py adds variable names in English to the catalog of raw files
    - 9_clean_yearbook.py creates datasets with yearbooks for all years
5. Prepares outcomes and merges all the data together
    - 10a_construct_variables_clevel reads county-level yearbooks into DTA file
    - 10b_construct_variables_prefabove reads prefecture-level yearbooks into DTA file
    - 11_clean_cityvar.do combines county- and pref-level data
    - 12_outcomes.do prepares a cleaned panel of city outcomes
    - 13_outlier_treatment.do drop outliers for the outcomes
    - 14_combine_ma_2016.do combine server simulations into one file, computed expected & recentered MA
    - 15_merge_data.do generate the main dataset analysis_data
6. Creates regression analyses and other outputs
    - Table1.do creates Table 1: Employment effects of market access
    - Table2.do creates Table 2: Regressions of MA growth on measures of economic geography
    - Maps_data.do outputs CSV data files for the maps
    - Misc_results.do produces other numbers directly reported in the draft

Helper files and Ado-packages:

code/ols_spatial_HAC.ado is a Stata program for Conley (1999) spatially clustered standard errors available, for instance, at http://www.trfetzer.com/conley-spatial-hac-errors-with-fixed-effects/.

Next, in *code/* we provide four original Stata ado-files that researchers may find useful in future applications. These ado-files are provided as-is and currently not available on the SSC archive; there are no proper helpfiles for them. The syntax is described at the top of each ado-file. Please email Kirill Borusyak if you have further questions or feedback. However, please do not email us about how to produce counterfactual shocks and recompute your treatment/instrument with them; this is entirely context-specific, and we are not able to help.

- *iv_spatial_hac*: implements Conley (1999) spatially clustered standard errors for instrumental variable regressions.

- Randomization inference for recentered IV and reduced-form regressions, after you have simulated many (e.g., 1,999) copies of your treatment or instrument for the same exposure but using counterfactual shocks
    - *ri_pvalue:* tests for beta=0 or beta=b for any other value of b in your recentered IV or reduced-form regression
    - *ri_ci:* computes the confidence interval for beta, i.e. the interval of all b which are not rejected at a given significance level (as in Table 1 of our paper)
    - *ri_spectest:* tests for correct specification of counterfactual shocks, by regressing the recentered instrument on a set of predetermined variables (as in Table 2 of our paper)

Instructions to Replicators

The entire data construction and analysis flow follows *code/master_hsr.do*. This do-file consists of six parts:

1. Setting directories and installing Stata packages
2. Preparing the railway database (in Stata)
3. Computing market access for actual and 1,999 counterfactual HSR maps, preferably on a research computing cluster
4. Extracting data from Chinese City Statistical Yearbooks (in Python)
5. Further data cleaning and merging (in Stata)
6. Regression and auxiliary analyses (in Stata)

To do complete the replication, the user should:

- Have working copies of Stata and Python
- Modify the path *$main* to the *BH replication* folder on your computer
- Have a stable internet connection; this will be used to install necessary Stata and Python packages
- Execute master_hsr.do up to the *stop* command
- Transfer the entire folder *server/* to a research computing cluster and start a job array of 100 parallel jobs of process_scenarios_2016.do. For clusters that use Torque or SLURM as job schedulers, we have provided the code to submit these jobs. Please update the *module load stata* line as necessary to load the correct Stata module on the cluster.
    - If the computing cluster uses Torque as a job scheduler, run *qsub run_server_2016_torque.sh*

    - If the computing cluster uses SLURM as a job scheduler, run *sbatch run_server_2016_slurm.sh*
    - If you have no access to a research computing cluster, you can execute this part without parallel jobs and without Linux, directly in Stata. The instructions are given in Part 3 of *master_hsr.do*. This may take ~70 hours.
- Once all jobs are complete transfer all files in *server/output* back into *$main/server/output*
- Make sure that *python* is in the system PATH. If your installation has a different command, e.g. *python3*, modify the calls in *master_hsr.do* accordingly.
- Execute the rest of *master_hsr.do*.

Maps (Figures 1, 2, and A1) were produced in ArcGIS ArcMAP 10.8.2 using the shapefiles listed above and CSV files output by *Maps_data.do* in Part 6 of *master_hsr.do*. The *png* files for the maps in our paper are available *gis/* in color and black&white versions:

- Figure 1.A: BH_lines_MA_2016*
- Figure 1.B: BH_lines_actual_planned*
- Figure 2.A: BH_expected_2016*
- Figure 2.B: BH_recentered_2016*
- Figure A1: BH_sim_lines_MA_2016*

To reproduce the maps, move the output files *lines_by_year.csv* and *ma2016.csv* into the *gis* folder after running the replication steps listed above. Opening the corresponding *.mxd* files in ArcMAP and exporting should yield the same map figures, without any extra steps.

List of tables and programs

The provided code reproduces all numbers provided in text, tables, and figures. The tables and figures are saved in the Results/ folder.

Figure/Table #	Program	Output file	Notes
Table 1	Table1.do	Table1.csv	
Table 2	Table2.do	Table2.csv	
Figure 1	ArcMap	BH_lines_MA_2016, BH_lines_actual_planned	
Figure 2	ArcMap	BH_expected_2016, BH_recentered_2016	

Figure A1 ArcMap BH_sim_lines_MA_2016

References

Brinkhoff, Thomas (2018). *City Population.* https://www.citypopulation.de/en/china/

China Railway Yearbook Editorial Board (2001-2013). *China Railway Yearbook*.

China Statistics Press (2000-2017). *China City Statistical Yearbook*.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. Journal of Econometrics 92:1–45.

Lawrence, Martha, Richard Bullock, and Ziming Liu (2019). China's High-Speed Rail Development. Washington, D.C.: World Bank.

Lin, Yatang (2017). "Travel costs and urban specialization patterns: Evidence from China's high speed railway system." Journal of Urban Economics 98:98–123.

OCHA Regional Office for Asia and the Pacific (2018). *Province and Prefecture Capitals of China.*

OCHA Regional Office for Asia and the Pacific (2020). *China - Subnational Administrative Boundaries.*