Data and Code README for

Commuting, Labor, and Housing Market Effects of Mass Transportation: Welfare and Identification

Christopher Severen
Federal Reserve Bank of Philadelphia

This README lists the data and code associated with the project "Commuting, Labor, and Housing Market Effects of Mass Transportation: Welfare and Identification," to be published at the *Review of Economics and Statistics.* One of the datasets I use cannot be shared: I use the Neighborhood Change Database; see below.

Unpacking

Because of prohibitively small maximum size restrictions for uploaded files to the ReStat Dataverse, I have broken up materials into several groups and compressed them (tar'd and sometimes gzip'd). These are:

- code.tar contains code, figures, results, tables foldes and master.do and profile.do files
- data.tar.gz contains main raw data files
- graphhopper.tar.gz contains graphhopper routing files
- output/output_main.tar.gz contains the main intermediate data files
- output/output_other.tar.gz contains additional intermediate data files

Organization of contents

In addition to this README file, this folder contains the following folders and files:

- ./code/ contains Stata and R scripts used to build the data used for analysis and perform and report results from the analysis.
- ./data/ contains raw input data used in the paper. I describe each of these data sources below.
- ./figures/ contains output figures.
- ./graphhopper/ contains the files and scripts used to simulate shortest-time travel routes, including the constructed java engine and OSM file inputs.
- ./output/ stores intermediate outputs, fully processed data, and other post-processing material that is not ultimately reported.
- ./results/ contains some preliminary model results and output.
- ./tables/ contains regression and model results.
- ./master.do is the main do file to execute the Stata portion of code that builds and analyzes the data.
- ./profile.do is automatically opened by master.do and sets useful settings.

See section below entitled "***Replication instructions***" for detailed instructions on how to replicate results and figures in the paper. This is followed by detailed information about all the files included in the folder.

This README has the following sections:

0. Computational requirements
1. Replication instructions
2. Data files (raw)
3. Data files (analysis)
4. Code and related files
5. Results

Obtaining Neighborhood Change Database data

To perform pre-trends analysis on tract-level covariates, I used Geolytics' Neighborhood Change Database (NCDB) licensed to the Federal Reserve Bank of Philadelphia. I include a more detailed description of the variables I use and provide processing code for those who have access to the data. To obtain access to the NCDB, interested researchers can contact:

Geolytics
P.O. Box 5336
Somerville, NJ 08876
T: 800-577-6717
E: questions@geolytics.com

0. Computational requirements

Computational Environment: All code was run on a PC running Windows 10 Enterprise (64-bit).

- Processor: Intel Core i7-8700 CPU @ 3.20 GHz
- 16.0 GB of RAM

Recommended Environment:

- Some of the data files held in memory by STATA are quite large, so I recommend at least 8 GB of RAM.

Required Software:

- STATA v16.0, with additional packages:
    - blindschemes
    - estout
    - reghdfe
    - ppmlhdfe
    - ftools
    - gtools
    - winsor2
    - stata-tex (from Paul Novosad: https://github.com/paulnov/stata-tex
- Python 3 or later (to execute stata-tex output)
- R v3.6 or later
- Java Virtual Machine (JVM)
- Grahhopper (routing that uses JRE: https://github.com/graphhopper/graphhopper/)

1. Replication instructions

Results from the last run of the analysis files are saved in the folder /results/. However, if you would like to replicate the results, you have two options:

Replication from intermediate data

This provides the quickest way to replicate data, as it takes intermediate data as given.

1. Open: / master.do , automatically executing ./profile.do
2. Execute lines 57-81 (some components will take a while to run). Make sure Dropbox or other syncing utilities are inactive when running, because the stata-tex utility sometimes edits files being actively written, throwing errors.
3. Open ./code/welfare/index_rwelfarescripts.R, set the appropriate <cdir> to the location of the ./, and execute those scripts to complete welfare analysis.

Replication from raw data

This takes some time and requires use of both R and Stata. To do thise:

1. Open: /code/build/index_rbuildscripts.R, set “cdir” appropriately, and execute.
2. Open: / master.do , automatically executing ./profile.do
3. Execute lines 19-48.
4. Follow the steps for “Replication from intermediate data” above.

2. Data files (raw)

The /data/ folders contains the following folders, each listed with contents, provenance, and details.

/data/CTPP/

Contains two folders (below) and /definitions.xlsx, a file I create to help concord variables across the two CTPP waves.

/data/CTPP/1990/

Includes data from the 1990 CTPP

- Can be accessed from https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=JFD&Yv0x=D
- File names refer to their series.
- I created /vars1990.xlsx to help track which variables are present.

/data/CTPP/2000/

Includes data from the 2000 CTPP

- Can be accessed from https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=JGD&Yv0x=D
- File names refer to their series.
- I create /ctpp2000_vars.xlsx to help track which variables are present

/Dynamap/

Travel time matrix simulated using a local variant of the Dynamap Road Network via ArcGIS.

/Geographies/

Groups of shapefiles from the National Historical Geographic Information System (NHGIS) used for geographic crosswalking.

- CA_blck_grp_2000 – Block Group shapfiles for California, 2000 vintage.
- CA_tract_1990 – Tract shapfiles for California, 1990 vintage.
- CA_tract_2000 – Tract shapfiles for California, 2000 vintage.
- CA_tract_2010 – Tract shapfiles for California, 2010 vintage.
- US_cty_sub_1990Consolidated – Subcounty shapefiles, national, 1990 vintage.
- Can be re-accessed by visiting https://www.nhgis.org/ and selecting variables and samples listed in the above codebook file.

/HousingDetails/

Contains tract and block-level statistics on residential population and housing characteristics for 1990 and 2000.

- Each of the ./nhgis00??_csv/ folders contains contains one or more pairs of .csv and .txt files:
    - Each .txt file is a codebook describing the variables and codes in each .csv file.
- The ./Housing Variables.xlsx and ./Housing Variables2.xlsx files I created to match variables across census years.

- Can be re-accessed by visiting https://www.nhgis.org/ and selecting variables and samples listed in the above codebook file.

/I105/

Shapefiles of major highways and roads in California. Each group of similarly-named files should be considered together, as this is how shapefiles are sometimes stored, but the principal file of each group is the one with the .shp extension.

- ./CA_AllMajor.* are from CalTrans, but I cannot find the specific file. However, these files are not used for any analysis
- ./CA_AllMajor_trim.* is a subset of CA_AllMajor.*, and is not used for any analysis.
- ./CA_NHS.* are from the National Highway System, set to just California roads and highways https://gisdata-caltrans.opendata.arcgis.com/datasets/8887c21dce1d45ad93cd78a10b24781b_0/about
- ./CA_NHS_trim.* are a subset of roads from CA_NHS.*, cut to Southern California.
- ./NHPNLine.* files are from the National Highway Performance Network: https://www.fhwa.dot.gov/planning/processes/tools/nhpn/index.cfm

/IPUMS/

Includes data from 5% samples of the 1980 and 1990 and 2000 Census and the 2006-2010 American Community Survey from IPUMS, used for constructing labor demand shocks.

- ./ipums_1980-0610_all.dta is the file produced by executing ./usa_00003.do on the downloaded dataset
- ./usa_00003.cbk.txt is the original codebook associated with the download.
- ./usa_00003.do is the do file from IPUMS that processed the original download file (no longer available) into ./ipums_1980-0610_all.dta
- Original download in *.dat format can be re-accessed by visiting https://usa.ipums.org/usa/ and selecting variables and samples listed in the above codebook file.

/KelkerDeleuw/

Hand-transcribed shapefiles of proposed and PER lines from Kelker, Deleuw, and Co. (1925) as described in the paper. As shapefiles, these are groups of files, but the *.shp files are the most important.

- ./Both1925Lines_notPER.* combine FutureConstr.*and ImmediateConstr.* and are used to create the Full 1925 Plan sample.
- ./FutureConstr.* are the lines meant to be built after the lines meant for immediate construction.
- ./ImmediateConstr.* are the lines meant for immediate construction and are used to create the Immediate 1925 Plan Sample.
- ./PER_All_LA_1925.* are the PER lines in Los Angeles in 1925, and are used to create the PER sample

/LEHD/

Origin-destination data for California in 2002 and 2015 from the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES), from LODES Version 7.

- ./ca_od_main_JT00_2002.csv contains the 2002 data
- ./ca_od_main_JT00_2015.csv contains the 2015 data
- Both files can be obtained from: https://lehd.ces.census.gov/data/

/Metro/

Shapefiles and supporting documentation from LA Metro's data site: https://developer.metro.net/gis-data/, along with additional supporting documentation I created to help evaluate counterfactuals.

- ./MetroROW/ contains right of way data from LA Metro.
- ./RailLines0614/ contains the 0614 version of rail line shapefiles.
- ./RailPortals0715/ contains the 0715 version of rail portal shapefiles.
- ./RailStations0715/ contains the 0715 version of rail station shapefiles.
- ./Stations0316/ contains the 0316 version of all station shapefiles.
- The other csv and xls files I create from information contained in the shapefiles or harvested from LA Metro's website.

/NCDB/

Empty folder where the appropriate Neighborhood Change Database files can be uploaded once obtained; see note on page 2.

/Oldmaps/

Copy of files in /KelkerDeleuw/

/Ridership/

Monthly ridership data per line from LA Metro that I edited into an easier-to-work-with format. The original file, which I no longer have, used to be downloadable on LA Metro's site. They now have a different interface for accessing statistics that does not allow queries before 2009, but custom data extracts can be requested by contacting ServicePerformanceAnalysis@metro.net.

/SCAG/

Zoning and land use data from the Southern California Association of Governments (SCAG) covering 1990, 1993, 2001, and 2005.

- ./SCAG_LU.* are a collection of shapefiles along with the layer package ./land_use_scag_2005.lpk that delineate zones of different types of land use in SCAG member counties for the four year listed above.
- ./Final_LU_class_02_update.doc contains land use codes and description.
- ./LandUse/ contains the land use for each of the polygons (tract interacted with land use code)in ./SCAG_LU.*, and can be used instead of opening the shapefiles.

/ThreeShocks/

This files contains details on three of the route alignment shocks mentioned in the paper: the Wilshire Boulevard corridor, the Green Line's western termination near LAX, and the Blue Line railroad alignment. While not used in the paper, these files were used in referee reports and I provide them here for posterity.

- ./DefiningThreeShocks.docx provides implementation notes and supporting evidence, as well as documentation for the following files.
    - ./GreenLine_LAXNorth_1992.pdfgives the LAX Green Line initial alignment.
    - ./HeavyRail1983_LinesStations.pdf gives the Red/Purple Line alignment along Wilshire Boulevard.
- ./UnbuiltLines.* contains shapefile information for these unbuilt lines that I constructed .
- ./UnbuiltStations.* contains shapefile information for the corresponding unbuilt stations that I constructed .

/TranspoDetails/

Contains tract and block group-level statistics on residential transportation characteristics for 1990 and 2000.

- Each of the ./nhgis00??_csv/ folders contains contains one or more pairs of .csv and .txt files:
    - Each .txt file is a codebook describing the variables and codes in each .csv file.
- Can be re-accessed by visiting https://www.nhgis.org/ and selecting variables and samples listed in the above codebook file.

/WageTransit/

Includes data from 5% samples of the 1980 and 1990 and 2000 Census and the 2008-2012 American Community Survey for Californian commuters aged 18-80 from IPUMS, used for studying heterogeneity of transit adoption by income.

- ./wagetransit.dta is the file produced by executing ./usa_00008.do on the downloaded dataset and saving it to that name.
- ./usa_00008.cbk.txt is the original codebook associated with the download.
- ./usa_00008.do is the do file from IPUMS that processed the original download file (./usa_00008.dat) into ./ipums_1980-0610_all.dta
- Original download in *.dat format can be re-accessed by visiting https://usa.ipums.org/usa/ and selecting variables and samples listed in the above codebook file.

3. Data files (analysis)

The /output/ folder and its subfolders contain data that has been processed. These are the files that are used for the analysis steps. We provide these as a convenience so that the data need not be re-constructed from scratch.

- ./bs_allparams_*.dta contain bootstrap estimates of parameters for the structural model, where * can be “all” for all data, “sim” for the immediate 1925 plan design, and “per” for the PER design.
- ./flows_final_lehd.dta contain processed LEHD LODES data for 2002 and 2015.
- ./flows_prepped_all.dta contains CTPP data in 1990 and 2000 and retains tract pairs for which in one or both years the flows are zero.
- ./flows_prepped_small.dta contains CTPP data in 1990 and 2000 and drops tract pairs for which in both years the flows are zero.
- ./ncdb_tractdata.dta contain NCDB data for pre-trends tests. **DATA ARE NOT INCLUDED IN PUBLIC RELEASE**
- ./powFEs.dta contain place-of-work fixed effects estimated from a gravity model
- ./routeproximity.dta contains information on how much of each fastest-origin-to-destination tract route is in each of several distance bins from the transit lines (used in congestion analysis)

/crosswalks/

Various files, created and used primarily in the build process, that link different geographies together.

/intermediate/

Various files, created and used in the build process, that store intermediate results of processing for ease of access.

/mapping/

Cleaned shapefiles used for mapmaking.

/routes/

Collections of polyline-containing shapefiles produced by graphhopper that represent the fastest routes from origin to destination centroids; used to create /output/routeproximity.dta

- I have OMITTED these from the main sharable extract due to size (all told about 40GB), but can provide them upon request.

/welfare/

Various data files used for welfare calculation.

4. Code and related files

All Stata-based analysis can be accessed by open /master.do. This also calls profile.do to set useful environmental variables. I describe the files used for each section of the paper below, however, some of the building and analysis also involves the use R, which is accessed through separate calls using RStudio. I describe those in more details where appropriate.

The /code/ folder contains:

- ./analysis/ contains code to perform the analysis, mostly relying on the intermediate data in /output/. I describe these in more details below.
- ./bartik-weight/ contains files to perform a shift-share decomposition from Goldsmith-Pinkham, Sorkin, and Swift (2020)
- ./build/ contains code to build the data, populating the intermediate data files in /output/. I refrain from describing this in details, but following master.do gives the executable order.
- ./tablecode/ contains Stata and Python scripts by Paul Novosad table construction (see "stata-tex")
- ./welfare/ contains R scripts to simulate model counterfactuals. I describe these files below.

Analysis files

/code/analysis/bootstrap_output.do

- Post-processes bootstrap data into a CSV file to be used by welfare analysis.

/code/analysis/bootstrap_run.do

- Executes bootstrapping procedure.
- Produces materials for Appendix Table E1.

/code/analysis/comparison_FEs_vs_wage.do

- Compares workplace fixed effects with average wages at workplace
- Produces materials for Table 3 Panel C.
- Produces Appendix Figure D1.

/code/analysis/epsilon_shiftshareanalysis.do

- Performs shift-share analysis after Goldsmith-Pinkham, Sorkin, and Swift (2020).
- Produces material for Appendix Table H7.

/code/analysis/epsilon_shiftshareanalysis_wsubcoFEs.do

- Performs shift-share analysis after Goldsmith-Pinkham, Sorkin, and Swift (2020), including subcounty-by-year fixed effects.
- Produces material for Appendix Table H8.

/code/analysis/estimates_lambdas.do

- Analyzes tract-level residuals and other relevant variables in 1990 and 200.
- Produces material for Table 5
- Produces material for Appendix Tables H9, H10, H11, and H12.

/code/analysis/finalflowcleaning.do

- Final cleaning and variable creation for flow data; called by other files.

/code/analysis/flows_congestion.do

- Creates congestion analysis.
- Produces materials for Table 2.

/code/analysis/flows_lehdthrough2015.do

- Performs LEHD LODES bilateral commuting analysis, 2002-2015.
- Produces materials for Table 1 Panel B.

/code/analysis/flows_metroeffects.do

- Primary analysis files for CTPP data (1990-2000).
- Produces materials for Table 1 Panel A.
- Produces materials for Appendix Tables H1, H3, H4, H5, and H6.
- Produces input materials for Appendix Figure H6 (see ../plot_treatmentgradient.do)

/code/analysis/gravity.do

- Performs and compares different methods for calculating gravity.
- Produces materials for Appendix Table F1.

/code/analysis/plot_flowdensity_ptreatment.do

- Plots probability of future treatment by 1990-level commuting flows.
- Produces Figure 3.

/code/analysis/plot_ridership.do

- Produces ridership plots shown in Appendix Figures H2 and H3.

/code/analysis/plot_ treatmentgradient.do

- Produces Appendix Figure H6.

/code/analysis/plot_ wagetransit.do

- Produces Figure 2.

/code/analysis/recover_powFEs.do

- Estimates place of work fixed effects used to estimate elasticities (see ../tracts_elasticities.do)

/code/analysis/test_ncdbpretrends.do

- Test pretrends over 1970-1990 using NCDB data.
- Produces materials for Appendix Table H2.

/code/analysis/tracts_elasticities.do

- Estimates tract elasticities and related calculations.
- Produces materials for Table 3 Panels A and B, and Table 4.
- Produces materials for Appendix Tables D1 and D2.

Welfare Files

/code/welfare/index_rwelfarescripts.R

- Indexes the other R scripts used in welfare analysis and preps environmental varaibles.

/code/welfare/prep_simdata.R

- Processes the data to be fed into the welfare model.
- (Most data is pulled from Stata build files).

/code/welfare/run_welfare_bootstrap.R

- Perform bootstrap estimation of model.
- Produces materials in Table 6.

/code/welfare/run_welfare_extended.R

- Simulate a variety of counterfactuals.
- Produces materials in Appendix Table H13.

/code/welfare/run_welfare_main.R

- Simulate main counterfactuals.
- Produces materials in Table 6.

/code/welfare/simcode_functions.R

- Functions that implement model calculations.

5. Results and Tables

The folder /results/ contains some initial results that are called later on. The folder /tables/ contains the set of tex and csv templates used by stata-tex to produce output tables and these output tables.. These are not quite in final table format; that step requires tweaking table format. Names are not consistent with the final version of the paper.

6. Graphhopper

The folder /graphhopper/ contains the materials used to spin up and query the graphhopper routing virtual machine. Note that Java is required.

After installing graphhopper execute the contents of the file ./call.txt in the terminal. This changes directories and runs initiates the virtual machine and routing engine using the relevant map data (socal-latest.osm.pbf) and configuration options (config-example.yml). The R script ./la_routingscript.R queries the engine and saves the results. This takes a long time to execute, ~1 week, and the resulting output is about 40GB.

TABLE CORRESPONDENCE AND RESULTS LOCATION

Table or Figure	**File that generates the results**
Figure 1	/code/build/map_prep3.R
Figure 2	/code/analysis/plot_ wagetransit.do
Figure 3	/code/analysis/plot_flowdensity_ptreatment.do
Table 1	/code/analysis/flows_metroeffects.do (Panel A) /code/analysis/flows_lehdthrough2015.do (Panel B)
Table 2	/code/analysis/flows_congestion.do
Table 3	/code/analysis/tracts_elasticities.do (Panels A and B) /code/analysis/comparison_FEs_vs_wage.do (Panel C)
Table 4	/code/analysis/tracts_elasticities.do (Panels A and B)
Table 5	/code/analysis/estimates_lambdas.do
Table 6	/code/welfare/run_welfare_main.R /code/welfare/run_welfare_bootstrap.R
Table A1	Not applicable; analytic.
Table D1	/code/analysis/tracts_elasticities.do
Table D2	/code/analysis/tracts_elasticities.do
Figure D1	/code/analysis/comparison_FEs_vs_wage.do
Table E1	/code/analysis/bootstrap_run.do
Table F1	/code/analysis/gravity.do
Figure H1	Not applicable; historical book pages.
Figure H2	/code/analysis/plot_ridership.do
Figure H3	/code/analysis/plot_ridership.do
Figure H4	Not applicable; descriptive table.
Figure H5	Not applicable; descriptive table.
Figure H6	/code/analysis/plot_ treatmentgradient.do
Table H1	/code/analysis/flows_metroeffects.do
Table H2	/code/analysis/test_ncdbpretrends.do
Table H3	/code/analysis/flows_metroeffects.do
Table H4	/code/analysis/flows_metroeffects.do
Table H5	/code/analysis/flows_metroeffects.do
Table H6	/code/analysis/flows_metroeffects.do
Table H7	/code/analysis/epsilon_shiftshareanalysis.do
Table H8	/code/analysis/epsilon_shiftshareanalysis_wsubcoFEs.do
Table H9	/code/analysis/estimates_lambdas.do
Table H10	/code/analysis/estimates_lambdas.do
Table H11	/code/analysis/estimates_lambdas.do
Table H12	/code/analysis/estimates_lambdas.do
Table H13	/code/welfare/run_welfare_extended.R