

Data for “Importing Polarization? The Electoral Consequences of Rising Trade Exposure”

David H. Autor
MIT and NBER

David Dorn
University of Zurich and CEPR

Gordon H. Hanson
UCSD and NBER

Kaveh Majlesi
Lund University and IZA

May 19, 2020

Main data sources

This paper studies outcomes that derive from four underlying data sources:

- The *Database on Ideology, Money in Politics, and Elections (DIME)* tabulates campaign contributions by donor and recipient for all amounts in excess of \$200, as mandated by the Federal Electoral Commission. It uses these contribution patterns across donors and candidates to infer measures of ideology. We use DIME version 3, which we obtained directly from its author, Adam Bonica of Stanford University. This version of the data comprises contributions until the 2016 Congressional Election, which is the last election studied in the paper. Note that future versions of the database that include campaign contributions for a longer time span may plausibly yield ideology estimates that are less representative for our period of analysis. DIME is freely available for download at <https://data.stanford.edu/dime>, although at the time of writing of this document, the published version of the database is version 2, which precedes the version we use.
- *Dave Leip's Atlas of U.S. Elections* is the source of our data on election outcomes for both congressional and presidential elections. It also provides information on registered voters. This data is available for purchase at <https://uselectionatlas.org/>.
- *Nielsen Local TV View* measures the size of the audience for a given programming hour of a given TV show on a given network. Our version of this data covers only major cable news channels, and is aggregated at the level of months, counties, and broad age-race/ethnicity groups. We obtained this data through an agreement between the National Bureau of Economic Research (NBER) and Nielsen Media. Further information on Nielsen data is available here: <https://www.nielsen.com/us/en/client-learning/tv/local/nltv/>.
- The *Pew Research Center Political Typology* (in 2004, 2011 and 2014) and the *Pew Research Center Governance Study* (in 2015) ask U.S. adults a consistent set of questions about their political beliefs. We follow Pew in using ten of these questions to construct a measure of ideology on a left-right scale. We obtained the PEW microdata under a data contract with the Pew Research Center. More information on Pew Data is available here: https://www.people-press.org/?post_type=dataset.

We do not publish the source data files, since the authors of these files have not made them freely and publicly available. The raw data will be preserved by us for ten years. Researchers who are interested in obtaining these data should consult the internet addresses indicated above.

Software environment (folder */do*)

All of the files described below are based on the statistical software Stata. We used Stata version MP 16.1, along with the downloadable Stata packages `ivreg2` (version 4.1.9), `reghdfe` (3.2.9), `ranktest` (1.4.1), `ftools` (2.37.0) and `unique` (1.2.3). The file *setup.do* installs the most recent updates of these packages.

Construction of analysis data (folder /do)

The file *consolidated_master_builder.do* executes the do-files *pew_master_builder.do*, *nielsen_master_builder.do*, *house_master_builder.do* and *president_master_builder.do*, which create all datafiles used for the regression analysis. These files document the data construction process, but they cannot be executed within this file archive since it does not include the raw data, as described above.

Analysis data files (folder /dta)

The online data archive contains the following analysis data sets in Stata format:

- *house_2002_2016.dta* provides data on 3,722 county-district cells. This file is the basis for all tabulated regression results that relate to congressional elections. Each cell is identified by a county and a district indicator (variables *cty_fips* and *congressionaldistrict*). The latter consistently refers to districts of the 108th Congress (elected in 2002), and we crosswalk all subsequent outcomes back to these initial geographic units. The main outcome variables are *d2_shnr_2002_Y* (the change in the Republican two-party vote share percentage between year 2002 and year $Y = 2004, \dots, 2016$), *d2_cfavg_X_2002_Y* (the change in the election probability for a candidate of a particular party-ideology combination where $X = \text{demlib, demmod, repmod, repcon}$ refers to liberal and moderate Democrats, and moderate and conservative Republicans, respectively), and *dhs2_cont_tcileT_2002_Y* (the proportionate change in campaign contributions from donors in ideology tercile $T = 1, 2, 3$, where the first tercile refers to the most liberal and the third to the most conservative donors). Local labor market exposure to import competition from China is measured by variable *d_imp_usch_pd*, which we instrument using the variable *d_imp_otch_lag_pd* whose construction draws Chinese exports to high-income nations other than the US. The control variables in specifications with full controls vary at the level of commuting zones (manufacturing employment share *l_shind_manuf_cbp*, occupation composition variables *l_sh_routine33* and *l_task_outsource*) or counties (lagged Republican vote shares in presidential elections *shnr_pres1996* and *shnr_pres2000*, Census division dummies *reg_**, and various population composition variables *l_sh_**).
- *descriptive_figures.dta* provides national time-series data on the ideological position of donors and elected candidates for congressional elections. The variable *cycle* indicates the year of the election. The variables *share_tot_cont_tcileT[_X]* indicate the fraction of campaign donations from donors in ideology tercile $T = 1, 2, 3$, where the first tercile refers to the most liberal and the third to the most conservative donors. By construction, the fraction of campaign contributions accounted for by liberal, moderate and conservative donors are $1/3$ each in 2002. Variables with suffix $X = \text{ind, corp, other}$ compute these shares only for the donations from individuals, corporations, or other types of donors. The variables *mean_w_dem_cfscore* and *mean_w_rep_cfscore* provide the average ideology score of elected Democratic and Republican

members of congress, where the underlying individual-level ideology scores of politicians are derived from the contribution-weighted average of the ideologies of their campaign donors.

- *president_2000_2016.dta* provides data for 3,107 counties. This file is the basis for regression results on presidential elections. The variables *shnr_presY* indicate the Republican two-party vote share for the elections of years $Y = 2000, 2008$ and 2016 . Other variables closely follow the naming conventions of file *house_2002_2016.dta*.

We cannot provide the data files for the analyses of Pew and Nielsen data.

Computation of results (folder */do*)

The following Stata do-files compute the results of the paper:

- *house_2002_2016_analysis.do* uses *house_2002_2016.dta* to compute the results for Tables 3 and 4, and Appendix Tables A1, A2, S2 (column 1), S5, S6, S8, S9, S10, S11, S12, and S13. Regression results are saved in log file */log/house_2002_2016_analysis.log*, and are manually tabulated in the Excel file */tab/ADHM-Tables.xlsx*.
- *regression_based_figures.do* uses *house_2002_2016.dta* to create Figures 4, 5, 6, and 7, and Appendix Figures A1 and S5. These figures are saved in the directory */gph*.
- *Descriptive_Figures.do* uses *descriptive_figures.dta* to create Figures 2, 3 and Appendix Figure S2. These figures are saved in the directory */gph*.
- *president_2000_2016_anaylsis.do* uses *president_2000_2016.dta* to compute the results for Table 5 and Appendix Table S2 (column 2). Regression results are saved in log file */log/president_2000_2016_analysis.log*, and are manually tabulated in the Excel file */tab/ADHM-Tables.xlsx*.
- *nielsen_2004_2016_analysis.do* computes the results for Table 2 and Appendix Tables S3, S4 and S7. Regression results are provided in log file */log/nielsen_2004_2016_analysis.log*, and are manually tabulated in the Excel file */tab/ADHM-Tables.xlsx*.
- *Nielsen_Graphs.do* create Figure 1 and Appendix Figures S1 and S4. These figures are saved in the directory */gph*.
- *pew_2004_2015_analysis.do* computes the results for Table 1 and Appendix Table S14. Regression results are provided in log file */log/pew_2004_2015_analysis.log*, and are manually tabulated in the Excel file */tab/ADHM-Tables.xlsx*.

The first four of these do-files are executable from this file archive. The latter three are not, because our data contracts prevent us from sharing the Nielsen and Pew data.

Log files, tables and figures (folders */log*, */tab* and */gph*)

The folder */log* comprises log files that are created by the Stata do-files above. The folder */tab* contains tabulations of descriptive statistics and regression results, both in a single Excel file and as individual PDF files. All tables are based on the outputs of the do-files above, except for Table S1 which lists questions from the codebooks for the Pew surveys. The folder */gph* provides PDF files of all figures. All of these figures created by the do-files above, except for Appendix Figure S3, which shows a map of North Carolina’s 12th district from Wikipedia that we overlaid with commuting zone boundaries.

Additional documentation (folder */other*)

We construct several crosswalks that may be of broader interest:

- *cw_cd108_cd109.dta*, *cw_cd108_cd110.dta*, *cw_cd108_cd113.dta* and *cw_cd108_cd115.dta* provide weighted crosswalks between congressional districts of the 108th Congress (elected in 2002) and congressional districts of the 109th, 110th, 113th and 115th Congress (elected in 2004, 2006, 2012, and 2016). The crosswalk for the 113th Congress covers all U.S. mainland states, while the other crosswalks cover just a small number of states that redistricted outside the usual decadal redistricting schedule (see Appendix Table A1 in the paper). We discuss our construction of these crosswalks in section 3.3 of the paper. In the file *cw_cd108_cd113.dta*, the variables *congressionaldistrict108* and *congressionaldistrict113* identify the districts of the 108th and 113th Congress, respectively. The variable *afact_cd108_cd113* indicates the fraction of the population of a given congress-108-district that maps to the geographic area of a given congress-113-district. It sums up to one for each district of the 108th congress. The population weights are based on the spatial distribution of population in the Census 2000. Our analysis maps outcomes of the 113th Congress to districts of the 108th Congress by weighting district-level outcomes of the 113th Congress with *afact_cd108_cd113* for each district of the 108th Congress. Researchers who instead want to map outcomes from the 108th Congress to the districts of the 113th Congress can correspondingly do so using the variable *afact_cd113_cd108*.
- *cw_ctycd108_ctycd109.dta*, *cw_ctycd108_ctycd110.dta*, *cw_ctycd108_ctycd113.dta* and *cw_ctycd108_ctycd115.dta* provide weighted crosswalks between cells of counties and congressional districts of the 108th Congress (elected in 2002) and cells of counties and congressional districts of the 109th, 110th, 113th and 115th Congress (elected in 2004, 2006, 2012, and 2016). In these files, county-district cells for congress X are identified by variable *ctycdX*. This seven-digit variable concatenates the 2-digit state FIPS code, the 3-digit FIPS county code within the state, and the 2-digit district code. County-district cells for the 108th, 109th and 110th Congress use county definitions from the year 2000 (variable *cty_fips_2000*), while the cells for the 113th and 115th Congress are based on county definitions for the year 2010 (variable

cty_fips_2010). The variable *affect_ctycd108_ctycd113* indicates the fraction of the population of a given county-congress-108-district cell that maps to the geographic area of a given county-congress-113-district cell, where population weights are based on the spatial distribution of population in the Census 2000. These crosswalks are only relevant for outcomes that are available at the level of county-district cells. Outcomes observed at the district level can be crosswalked with the district-to-district crosswalks described above.

For additional related data, see David Dorn's data page, www.ddorn.net/data.htm.