Data for "Foreign Competition and Domestic Innovation: Evidence from U.S. Patents"

David Autor
MIT and NBER

David Dorn
University of Zurich and CEPR

Gordon H. Hanson
UC San Diego and NBER

Gary Pisano
Harvard University

Pian Shu
Georgia Institute of Technology

January 22, 2019

Data files (folder */dta*)

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

- *firm_7507_main_data_final.dta* provides firm-level data for four time periods (1975-1983, 1983-1991, 1991-1999, 1999-2007). The variable *yr* indicates the start year of a period. All variables with prefix "*dhs_*" are outcome variables defined as the change in an outcome between t and t+1, divided by the average value of the outcome in t and t+1. The variable *d_import_usch_pd* is the growth of Chinese import penetration in a firm's main industry, and *d_import_otch_pd_lag* is an instrument for this variable that is constructed using third countries' imports from China. The variables *ntrgap* and *mfa_schott* are the tariff gap and MFA quota instruments constructed by Pierce and Schott (2016). Control variables for the regression analysis include a set of dummy variables for manufacturing sectors (*mfg_ **), start-of-period characteristics of industries (*l_ind_ **), an indicator for U.S.-headquarterd firms (*firm_us*), a firm's log U.S. sales and its R&D-to-sales ratio (*ln_sales_us, xrd_sale*), and the fraction of a firm's patents by broad technology category (*lx_sh_tcat_ **).

- *firm_9107_hist_data_final.dta* provides firm-level data for two time periods (1991-1999, 1999-2007). It computes firms' import exposure based on historical industry affiliations, and includes an indicator variable for firms that change their main industry during a period (*indchg*).

- *firm_9107_segm_data_final.dta* provides firm-level data for two time periods (1991-1999, 1999-2007). It computes firms' import exposure based on historical information about their activity across multiple industry segments, and includes indicator variables for firms that enter or exit an industry segment during a period (*seg_add, seg_drop*).

- *firm_9107_nopat_data_final.dta* provides firm-level data for two time periods (1991-1999, 1999-2007). It includes information from firms that were observed in Compustat both at the start and end of a period, but which did not submit a successful patent application in either year.

- *tech_9107_main_data_final.dta* provides data at the level of patent technology classes for two time periods (1991-1999, 1999-2007). The outcome variables *dhs_patent_ ** measure the change in patenting either for Compustat corporations (suffix **_cscorp*), for all corporations (**_allcorp*), for all non-corporations (**_noncorp*), or for corporations and non-corporations jointly (**_allentities*).

- *pat5yr_adhps.dta* provides patent-level data for five application years (1975, 1983, 1991, 1999, 2007). It includes patents with corporate assigneee and U.S.-based main inventor. The variables *appyear* and *gyear* indicate application and grant years, respectively. The variables *matched_ADHPS* and *match_source* indicate whether a patent is linked to Compustat firms using our crosswalk and if so, the matching method. The variable *webmatch* indicates whether our web search algorithm finds a match for a patent. The variable *matched_pdp* indicates

whether a patent is linked to Compustat in the NBER Patent Data Project. The dataset also includes variables on patents' sectoral affiliations (*mfg, mfg_ *, nmfg, sector*) for patents matched to Compustat firms with valid industry affiliations (*valid_ sic*), as well as variables indicating whether a patent belongs to the various samples used in our analyses (*sample_ **).

- *pat5yr_ adhps_ foreign.dta* and *pat5yr_ adhps_ noncorp.dta* provide additional patent-level data for five application years (1975, 1983, 1991, 1999, 2007). These files include patents by foreign-based inventors and by non-corporate assignees, respectively.

- *figure1a_ pat_ imports.dta, figure1b_ imports.dta, tableA1_ IBM.dta,* and *tableA3_ compustatinfo.dta* comprise aggregate data used to construct the corresponding figures and tables.

Computation of results (folder */do*)

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

- *firm_ 7507_ main_ analysis_ final.do* computes all tabulated regression results in Tables 1-4 and Appendix Tables 5-9.

- *summarystats.do* creates Figures 1a, 1b, and Appendix Tables A1-A4. It also computes the patent counts indicated in Appendix Tables A5-A8.

Raw results (folders */log* and */tabfig*)

The folder */log* contains the Stata log files created by *firm_ 7507_ main_ analysis_ final.do* and *summarystats.do*, which contain the raw results for all tables except for Appendix Tables A1 and A4. The folder */tabfig* contains additional results created by *summarystats.do*: Raw tabulations for Appendix Tables A1 and A4 and PDF files of Figures 1a and 1b.

Assignment of patents to Compustat firms (folder */patent-match*)

The folder */patent-match* provides additional documentation for our matching between patent data and Compustat records:

- *cw_ patent_ compustat_ adhps.dta* provides a crosswalk between patent numbers from USPTO (variable *patent*) and company identification keys from Compustat (variable *gvkey*). It includes all utility patents with corporate assignees (variable *corpasg*=1) that were granted between January 1975 and March 2013 and are matched to Compustat. The variables *appyear* and *gyear* indicate application and grant years, respectively. The variable *usinv* indicates whether the primary inventor of a patent is based in the US.

- *Readme Patent Match.pdf* provides a step-by-step description of our patent matching algorithm, and explains the use of the auxiliary files that are provided in the subfolders *patent-match/python* and *patent-match/stata*.