DISCERN, August 2020

By Ashish Arora, Sharon Belenzon, and Lia Sheer

Main Data Sources

Due to intellectual property (IP) restrictions, not all data can be redistributed. In such cases, we provide (a) the program code as a reference for users, and (b) aggregated outputs at the firm-year level (when possible). Users who wish to obtain the data and use the program code are welcome to contact us directly with any questions.

(1) Company and accounting information from Standard & Poor's Compustat were obtained through Wharton Research Data Services (WRDS) in August 2018. Users should obtain the North American Compustat data file.

(2) Scientific publications were obtained from Clarivate Analytics' Web of Science (WoS). We include articles from journals covered in the "Science Citation Index" and "Conference Proceedings Citation Index - Science", while excluding social sciences, arts, and humanities articles. The publication data in the replication package are only available at the aggregate "permno_adj-year" level due to IP restrictions.

(3) U.S. Granted Patents for 1980-2015 were obtained from the U.S. Patent and Trademark Office (USPTO). The main patent file with ownership information at grant year is available under "output_files/DISCERN_patent_database_1980_2015_final1.dta".

(4) Related Non-Patent Literature (NPL) citations and forward patent citations were obtained from the 2016 edition of PATSTAT, the European Patent Office's Worldwide Patent Statistical Database. The patent citation data in the replication package are only available at the aggregate "permno_adj-year" level due to IP restrictions.

(5) Subsidiary data were obtained from (a) historical snapshots of Bureau van Dijk's Orbis Ownership files for 2002-2015, and (b) the NBER patent database for pre-2002 ownership data (see Hall, Jaffe, and Trajtenberg, 2001 and Bessen, 2009). The Orbis database does not allow data redistribution.

(6) Mergers and acquisitions data for 1980-2015 were obtained from Refinitiv's Securities Data Company (SDC) Platinum. This database does not allow data redistribution.

(7) Company name changes were obtained from WRDS's CRSP Monthly Stock. This database does not allow data redistribution.

For (5), (6), and (7), we provide the output of our data work for users to match to their database of interest. The ultimate owner (UO) and subsidiary historical standardized name lists ("output_files/DISCERN_UO_name_list.dta" and "output_files/DISCERN_SUB_name_list.dta", respectively) include the dynamic reassignment of firms. More information on the data construction process can be found in the Online Data Appendix.

Folder Structure

The contents of the replication data folders are summarized below.

Data folder	Content	Note
data	Raw and intermediate data files	See the description of files in the next section.

Data folder	Content	Note
programs	Text files containing Stata program commands	The do-file named "main_do_file.do" is the master file that runs all other do-files in the "programs" folder.
output_files	Main output files	Includes the main accounting panel data file with basic aggregated variables ("DISCERN_Panel_Data_1980_2015.dta"; uses "permno_adj_long" as unique identifier), the patent level output file ("DISCERN_patent_database_1980_2015_final1.dta"), the historical name lists with dynamic reassignment ("DISCERN_SUB_name_list.dta" and "DISCERN_UO_name_list.dta").

Dataset List

The sources of the raw and intermediate data files are summarized below.

Data file	Source	Note	Provided
data/corp_NPL_output_matched_short.dta	PATSTAT, WoS	Raw cleaned NPL match to publication; we cannot provide due to IP restrictions of PATSTAT and WoS	No
data/corp_NPL_cite_per_year_firm_80_15.dta	PATSTAT	Aggregated at "permno_adj-year" level	Yes
data/dyn_match_All.dta	Orbis, SDC Platinum, CRSP Monthly Stock, NBER	Dynamic match between firm names and ultimate owners; combines multiple data sources	Yes
data/fillin_gap_years.dta	Manual	Used for panel construction; see "panel_do.do" for more details	Yes
data/pat_per_year_permno_adj.dta	USPTO	Aggregated patent flow per permno_adj-year	Yes
data/pat_stock_permno_adj.dta	USPTO	Aggregated patent stock, including reassignment, per permno_adj-year	Yes
data/patent_1980_2015.dta	USPTO	Intermediate patent construction file	Yes
data/patent_firms.dta	Compustat, USPTO	Patent firm list	Yes

Data file	Source	Note	Provided
data/patent_match_id_name.dta	Compustat, USPTO, PATSTAT	Raw cleaned patent match to names	Yes
data/permno_gvkey.dta	Compustat	The connection to the North American Compustat file is made through "gvkey" codes; this file contains the link between "permno_adj" and "gvkey" codes	Yes
data/permno_min_max_year_adj_80_15.dta	Compustat	Intermediate file with range of years for each permno_adj firm	Yes
data/pub_match_id_name.dta	Wos	Raw cleaned publication match to names; we cannot provide due to IP restrictions of WoS	No
data/pub_per_year_permno_adj.dta	WoS	Aggregated publication flow per permno_adj-year	Yes
data/pub_stock_permno_adj.dta	WoS	Aggregated publication stock, including reassignment, per permno_adj-year	Yes
data/bdc2d66c4378743b.dta	Compustat	Company and accounting information from Compustat; We cannot provide due to IP restrictions	No

Computational Requirements

The Stata code for data cleaning and analysis was last run using StataMP15 (64-bit) on a personal laptop using Windows 10-Pro with 24GB RAM.

Instructions

(1) Users should read the Online Data Appendix before using the data.

(2) Users should refer to the "main_do_file.do" file available in the "programs" folder to replicate the data cleaning and analysis steps. This do-file makes the connection to all other do-files in the "programs" folder.

Description of Programs

(1) The file "programs/main_do_file.do" runs all other do-files files in the "programs" folder.

(2) The file "programs/compustat_do.do" compiles financial data based on North American Compustat records.

(3) The file "programs/patent_do.do" compiles patent data, including flow and stock variables with dynamic reassignment of patents.

(4) The file "programs/pub_do.do" compiles publication data, including flow and stock variables with dynamic reassignment of publications.

(5) The file "programs/npl_do.do" compiles NPL citation data.

(6) The file "programs/panel_do.do" compiles the accounting panel data.

(7) The file "programs/NPL_cleaning_exp.do" provides sample code for cleaning NPL citations.

(8) The file "programs/NAME_STD.do" provides sample code for standardizing the name lists.

References

Bessen, James. 2009. "NBER PDP Project user documentation: Matching patent data to Compustat firms." Unpublished documentation. URL: http://users.nber.org/~jbessen/matchdoc.pdf

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg. 2001. "The NBER patent citation data file: Lessons, insights and methodological tools." NBER Working Paper 8498, National Bureau of Economic Research. URL: https://www.nber.org/papers/w8498