DISCERN Datasen

By Ashish Arora, Sharon Belenzon, and Lia Sheer

December 2020

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 North America Annual Compustat (1980-2015). Available to Duke University through Wharton Research Data Services (WRDS) by navigating to "Home > Get Data > Compustat - Capital IQ > Other Compustat > North America - Annual Updates > Compustat Annual Updates - Fundamentals Annual" from the page https://wrds-web.wharton.upenn.edu/wrds/ds/comp/funda/index.cfm?navId=80 (accessed August 2018). Due to intellectual property (IP) restrictions, users should obtain the North America Annual Compustat data file.

(2) Scientific publications were obtained from Web of Science (WoS) Core Collection from Clarivate Analytics. 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 WoS Core Collection XML dataset covering 1900-2016 was obtained from Clarivate Analytics by license to Professor Sharon Belenzon and his research team (August 2017). Pricing: cannot disclose due to contract confidentiality clause. The files included: (i) Science Citation Index Expanded XML (SCIE XML) – 1900-2016; (ii) Conference Proceedings Citation Index-Science & Technical XML (CPCI-S XML) – 1990-2016; (iii) Times Cited file; (iv) PubMed ID (PMID) crosswalk file – bridge file between WOSID and PMID. Contact for subscription and pricing: Tim Otto | Clarivate Analytics (US) LLC, 1500 Spring Garden Street, Fourth Floor, Philadelphia, PA 19130, USA| Email: timothy.otto@clarivate.com | Phone: +1 215-490-5661. The publication data in the replication package are only available at the aggregate "permno_adj-year" level due to IP restrictions. Users can obtain publication data either by purchasing a license to WoS or by downloading data from available open sources.

(3) U.S. Granted Patents for 1980-2015 from the U.S. Patent and Trademark Office (USPTO) were obtained through PATSTAT Global – single edition (2016), The European Patent Office's Worldwide Patent Statistical Database. A license was purchased by Professor Sharon Belenzon in 2016 for approximately 1,000 EUR. To purchase a license, users should visit https://www.epo.org/searching-for-patents/business/patstat.html. The matched patent file is available under "output_files/DISCERN_patent_database_1980_2015_final1.dta". Users can obtain patent information either by purchasing a license from PATSTAT or by downloading data from available open sources.

(4) Related Non-Patent Literature (NPL) citations and forward patent citations were obtained through PATSTAT Global – single edition (2016), The European Patent Office's Worldwide Patent Statistical Database. A license was purchased by Professor Sharon Belenzon in 2016 for approximately 1,000 EUR. To purchase a license, users should visit https://www.epo.org/searching-for-patents/business/patstat.html. The patent citation data in the replication package are only available at the aggregate "permno_adj-year" level due to IP restrictions. Users can obtain NPL data either by purchasing a license to PATSTAT or by downloading data from available open sources.

(5) Subsidiary data were obtained from:

(a) historical snapshots from Orbis Ownership files for 2002-2015 from Bureau van Dijk. The Orbis database does not allow data redistribution. Source: via a Duke University subscription to Bureau van

Dijk's Orbis, Osiris, and Amadeus databases. Pricing: cannot disclose due to contract confidentiality clause. Contact for subscription and pricing: Jordan Hoffa, Business Development Manager, Bureau van Dijk, 120 North LaSalle, Chicago, IL 60602, USA | Email: jordan.hoffa@bvdinfo.com | Phone: +1 312 235 2515.

(b) the NBER patent database for pre-2002 ownership data (see Hall, Jaffe, and Trajtenberg, 2001 and Bessen, 2009). NBER 2001 data are freely available at: http://data.nber.org/patents/, while NBER 2006 data are freely available at: https://sites.google.com/site/patentdataproject/Home/downloads?authuser=0.

(6) Mergers and acquisitions data for 1980-2015 were obtained from Securities Data Company (SDC) Platinum by Refinitiv. This database does not allow data redistribution. Source: via a Duke University subscription to SDC Platinum - Mergers & Acquisitions module. Pricing: cannot disclose due to contract confidentiality clause. Contact for subscription and pricing: Crystal Muntz, Account Manager, Refinitiv | Email: crystal.muntz@refinitiv.com | Phone: +1 314-468-2004.

(7) Company name changes were obtained from WRDS's CRSP Stock (Monthly) and CRSP Compustat Merged (Monthly). This database does not allow data redistribution. CRSP Stock (Monthly)is available to Duke University through WRDS by navigating to: "Home > Get Data > CRSP > Monthly Update > Stock > Security Files > CRSP Monthly Stock" from the page https://wrds-web.wharton.upenn.edu/wrds/ds/crsp/stock_m/msf.cfm?navId=145. CRSP Compustat Merged (Monthly) is available to Duke University through WRDS by navigating to: "Home > Get Data > CRSP > Monthly Update > CRSP > Compustat Merged Database - Linking Table" from the page https://wrds-web.wharton.upenn.edu/wrds/ds/crsp/ccm_m/linktable/index.cfm?navId=137.

For (5), (6), and (7), we provide an aggregate 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 the files in the next section.
programs	Text files containing Stata and Python program commands	The do-file named "main_do_file.do" is the master file that connects 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	PATSTAT /USPTO	Aggregated patent flow per permno_adj-year	Yes
data/pat_stock_permno_adj.dta	PATSTAT /USPTO	Aggregated patent stock, including reassignment, per permno_adj-year	Yes
data/patent_1980_2015.dta	PATSTAT/USPTO	Intermediate patent construction file	Yes
data/patent_firms.dta	WRDS Compustat, PATSTAT/USPTO	Patent firm list	Yes
data/patent_match_id_name.dta	WRDS Compustat, PATSTAT/USPTO	Raw cleaned patent match to names	Yes
data/permno_gvkey.dta	WRDS 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	WRDS Compustat	Intermediate file with a range of years for each permno_adj firm	Yes
data/pub_match_id_name.dta	WoS	Raw cleaned publication match to names; we cannot	No

Data file	Source	Note	Provided
		provide due to IP restrictions of WoS	
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	WRDS Compustat	We cannot provide due to IP restrictions	No

Computational Requirements

The sample is at the firm-year level and includes an unbalanced panel of 60,885 observations. 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. All replication programs included have a run time of less than 1 hour.

Instructions

(1) Users should read the Online Data Appendix before using the data. Make sure you download the most updated version of 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. Users should be aware that some parts of the code cannot be replicated due to IP restriction on the data. In such cases, the code is available for reference.

(3) Users can find a list of all the variables, alongside their descriptive labels, under "programs/ Panel_do.do".

Description of Programs

(1) The file "programs/main_do_file.do" connects to 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. http://users.nber.org/~jbessen/matchdoc.pdf. Data available at: https://sites.google.com/site/patentdataproject/Home/downloads?authuser=0 (Accessed: April, 2016).

Bureau van Dijk. 2018. ORBIS Ownership files, 2002-2015. Bureau van Dijk, Chicago, IL. Provided via a Duke University subscription service.

Center for Research in Security Prices (CRSP). 2018. CRSP Compustat Merged (Monthly), 1980-2015. Available to Duke University through Wharton Research Data Services (WRDS). https://wrds-web.wharton.upenn.edu/wrds/ds/crsp/ccm_m/linktable/index.cfm?navId=137 (Accessed: August 2018).

Center for Research in Security Prices (CRSP). 2018. CRSP Stock (Monthly), 1980-2015. Available to Duke University through Wharton Research Data Services (WRDS). https://wrds-web.wharton.upenn.edu/wrds/ds/crsp/stock_m/msf.cfm?navId=145 (Accessed: August 2018).

Clarivate Analytics. 2016. Web of Science (WoS) Core Collection XML, 1900-2016. Clarivate Analytics, Philadelphia, PA. Obtained from Clarivate Analytics by license in 2017.

European Patent Office. 2016. The EPO Worldwide Patent Statistical Database, 1980-2015. PATSTAT Global single edition 2016. Obtained from EPO by license in 2016.

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. https://www.nber.org/papers/w8498. Data available at: http://data.nber.org/patents/ (Accessed: April 2016).

Securities Data Company (SDC) Platinum. 2018. Mergers & Acquisitions module, 1980-2015. Refinitiv. Provided via a Duke University subscription service.

Standard & Poor's (S&P). 2018. Compustat Segments, 1980-2015. Available to Duke University through Wharton Research Data Services (WRDS). https://wrds-web.wharton.upenn.edu/wrds/ds/comp/seghistd/index.cfm?navId=87 (Accessed: August 2018).

Standard & Poor's (S&P). 2018. North America Annual Compustat, 1980-2015. Available to Duke University through Wharton Research Data Services (WRDS). https://wrds-web.wharton.upenn.edu/wrds/ds/comp/funda/index.cfm?navId=80 (Accessed: August 2018).