**Matching USPTO Patent Assignees to Compustat Firms**

Pian Shu

January, 2019

Autor, Dorn, Hanson, Pisano, and Shu (2019) match USPTO patents to Compustat firms. The Online Appendix of the paper details the matching process and presents various summary statistics.

This document presents a step-by-step guide to matching USPTO patent assignees to Compustat firms. The supplementary Stata and Python programs in subfolders "/stata" and "/python" contain code for the key steps of this matching process.

Step 1: Download the source data:

- Patent data from the U.S. Patent and Inventor Database (the March 2013 version is used here).
- Linked patent-Compustat data from the NBER Patent Data Project, which covers only patents granted by 2006.
- Compustat North America (typically via institutional access). The relevant variables include Compustat firm ID (gvkey), firm name, firm website, current and historical SIC codes, financial statement data (e.g., sales and R&D expenditure), and segment sales data (i.e., sales by industry).

Step 2: Clean assignee names and Compustat firm names.

a. The first step in name cleaning is to remove punctuation and accent marks. The do files "clean_assignee" and "clean_compustat" contain the code for removing punctuation in the assignee and Compustat firm names respectively. The punctuation-free names are used as input into the Bing search; they are also used in string matching.
b. The second step in name cleaning is to apply the name-standardization procedures of NBER PDP (i.e., Derwent standardization). Doing so entails standardizing entity names (e.g., converting "INCORPORATION" and "INCORPORATED" to "INC") and condensing abbreviations (e.g., converting "A B C INC" to "ABC INC"). The do file "clean_assignee" contains the code for applying Derwent standardization to patent assignee names, which can be easily modified for Compustat firms.
c. The do file "clean_assignee" also contains the code from NBER PDP for identifying types of assignees (e.g., corporation, university, or government) based on their names.

Step 3: Use the Bing Web Search API to collect search results in the form of URLs.

a. Create the csv input file that contains two variables: (i) firm ID, generated using the punctuation-free name, and (2) the punctuation-free firm name.
b. Run the Python program "bing_asgsearch.py" after adding the API key to "bing_searchweb.py." A paid subscription to the Bing Web Search API is required when performing more than 3000 searches in a month.

c. The Python programs will generate an output file in csv that contains the links, titles, and descriptions of the top five search results from searching the punctuation-free firm names in quotation marks on Bing.
d. The do file "clean_url" contains the code for cleaning the URLs collected from the Bing search, e.g., for removing common prefixes such as "http://" and suffixes such as "index.html." It also identifies and drops aggregate websites such as yellow pages, which may show up as top search results for multiple firms.

Step 4: Match assignees to Compustat firms using names and URLs.

- We consider a patent assignee and a Compustat firm to be a match if one of the following conditions is met (in descending order of matching strength):
    i. The punctuation-free names match exactly.
    ii. The Derwent-standardized names match exactly.
    iii. One of the top five search results for the assignee exactly matches the firm website listed in Compustat.
    iv. The top five search results for the assignee and for the Compustat firm share at least two exact matches.
    v. The punctuation-free names share the same first word, and the firm website listed in Compustat is a partial string of one of the top five search results for the assignee (or vice-versa).
- Within each tier of matching, it is possible (but very rare) that a patent is matched to multiple gvkeys. We apply tiebreakers based on the availability of segment sales data, historical industry affiliation, and R&D spending data.
- For patents unmatched via name and web match, we add the matching from NBER PDP and manual search, if available:
    i. We use the Derwent-standardized assignee names to extend the matching in NBER PDP to patents granted after 2006. When available, we match patents unmatched in Step 4 to gvkeys from (the extended) NBER PDP.
    ii. For the remaining unmatched patents, we identify the 200 largest assignees (i.e., those with the most unmatched patents). We then manually match them to Compustat firms (when a match can be found).

The file "cw_patent_compustat_adhps.dta" comprises the crosswalk between USPTO patents and Compustat used in Autor, Dorn, Hanson, Pisano, and Shu (2019). The dataset was constructed in April 2016 and contains USPTO utility patents granted by March 2013.

**Reference:**

Autor, David, David Dorn, Gordon H. Hanson, Gary Pisano, and Pian Shu, "Foreign Competition and Domestic Innovation: Evidence from U.S. Patents," *American Economic Review: Insights*, forthcoming.