NBER PDP Project

User Documentation:

Matching Patent Data to Compustat Firms

Version: Beta, May 2009

by James Bessen

1 Overview

Researchers often want to know information about the patents owned by firms or the characteristics of patent owners. We provide information to facilitate matching PDP patent data to Compustat data maintained by Wharton Research Data Services (WRDS).

This task is more difficult than it might initially seem because we only have limited information about ownership and because Compustat data are about securities, not firms *per se*. It is helpful to explain what we mean by "own" and "firm" and how we developed our matching data.

Our initial data about ownership comes from Patent Office data, specifically, the name of the organization(s) to which the patent was assigned at issue.[1] Most of our match data come from procedures that match the initial assignee names to lists of corporate names and their subsidiaries.

We performed this matching through a multi-step procedure. We began with the matches identified in the 1999 NBER patent data, which were based on matching assignee names to firms and subsidiaries identified in *Who Own's Whom?* We then used an automatic name matching routine that began by cleaning and standardizing names. This routine removed designators of corporate form (e.g., "Inc.") and standardized common abbreviations. Using these cleaned names, we were able to identify a large number of cases where the standardized assignee name exactly matched a standardized organization name.

We then used a word frequency algorithm to identify likely matches. This routine generated a score for each potential match, the score based on the inverse of the frequency of each word in the name. Potential matches that included unusual words in both the assignee name and the organization name received high scores. High-scoring matches were then examined manually, sometimes checking corporate databases for additional information to verify whether they did, in fact, match.

This procedure provides us with a large number of assignee-organization matches. These organizations, however, do not necessarily correspond directly to records within the Compustat data for several reasons. First, Compustat records identified by CUSIPs or GVKEYs refer to securities, not firms. A single organization may correspond to multiple entries within the Compustat data. Sometimes reorganization of the ownership structure generates a new GVKEY; sometimes accounting changes result in multiple GVKEYs for the same organization; sometimes a parent organization and a subsidiary will both have GVKEYs. In order to uniquely identify organizations, we introduce a variable named PDPCO. In most cases, the PDPCO equals the Compustat GVKEY, however, in some cases, multiple GVKEYs are associated with a single PDPCO. These associations are recorded in the PDPCOHDR file.

Second, although the initial assignees are the initial owners of patents, sometimes ownership changes (and not only for corporate reorganizations). Ownership of the organization changes through mergers, acquisitions, spinoffs, etc. We assume that when an organization is acquired/merged/spun-off that its patents go to the new owner. We use data on mergers and acquisitions of public companies reported in the SDC database to track these changes.

Patent ownership also changes when patents are re-assigned, that is, sold independently of their assigned organization. At present we do not track these re-assignments.

Since owners change over time, patents must be matched to owners dynamically. We record these dynamic matches in the DYNASS file. For each initial assignee, it reports up to five corporate owners. The following describes the PDPCOHDR and DYNASS files and their uses.

---

1 About 5% of patents are assigned jointly to multiple organizations.

In addition to data on matches, we also obtained data on definite *non*-matches, that is, Compustat firms that do not appear to have any patents. We identified these by using a word matching routine. First, we identified words (longer than four characters) that were unique within all of the firm names in Compustat (e.g., "Primerica"). If that word was not found among all the words in the names of patent assignees, then we classified this firm as a non-match.

Finally, it is also worth noting what we mean by a patent "assignee." Assignee names are listed in the patent data from the USPTO, but they are not standardized.[2] For example, there are over 100 different spellings, misspellings, abbreviations, etc. for the assignees of patents assigned to IBM. Using extensive name standardization and matching routines, we grouped these into a single "assignee" that is assigned a unique number stored in the variable PDPASS. This number is based on the first patent the firm was granted after January 1, 1976. Note that some firms do not have utility patents, but do have other patents, e.g., design patents. These firms have negative PDPASS numbers.

2 The USPTO puts out a PATSIC product that includes standardized names. We incorporated their standardization in ours, but we did more extensive matching and we also assigned a unique assignee number that will not change over time.

2 Company Header File

The contents of the PDPCOHDR file are as follows:

```
Contains data from pdpcohdr.dta
  obs:        25,480                          Compustat 2006 firms assigned
                                                to PDP company and matched to
                                                patents v. alpha
 vars:            10                          7 Apr 2008 09:02
 size:     1,503,320 (97.8% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
name            str28  %28s                   Company Name
cusip           str6   %6s                    CUSIP Issuer Code
firstyr         int    %9.0g                  First year sales is good
gvkey           long   %9.0g                  SPC Permanent Number
lastyr          int    %9.0g                  Last year sales is good
pdpco           long   %9.0g                  PDP company
pdpseq          byte   %9.0g                  GVKEY sequence within PDPCO
begyr           int    %9.0g                  Beginning year for GVKEY within
                                                PDPCO
endyr           int    %9.0g                  Last year for GVKEY within PDPCO
match           float  %9.0g                  Matched to patents
-------------------------------------------------------------------------------
Sorted by:  gvkey
```

The variables are:

CUSIP	The Standard and Poor's security identifier. Note that this does not uniquely identify a company, since firm securities change and Cusip numbers are re-used.
NAME	Company security name
FIRSTYR	First year GVKEY company has data
GVKEY	GVKEY is WRDS identifier of Compustat records; this is consistent over time (as Cusip numbers change), but it changes with corporate reorganizations
LASTYR	Last year that GVKEY company has data
PDPCO	Unique identifier of company; this may include multiple GVKEYs if company is reorganized or if multiple records are listed in Compustat
PDPSEQ	Values = 1-4: these value designate the sequence of the GVKEY within the PDPCO (see discussion below)
	Value = -1: the data for this GVKEY are not used for this PDPCO
BEGYR	First year data for this GVKEY are used for the PDPCO

ENDYR	Last year data for this GVKEY are used for the PDPCO
MATCH	Value = 1 : the PDPCO is matched to a patent assignee (although the individual GVKEY may not be matched)
	Value = 0 : the PDPCO is a definite non-match
	Value = Missing : Match/non-match not known

When PDPCOs include multiple records, we assign a single GVKEY record to by used for each year there is data. For example, consider Celanese Corp., which went through several reorganizations:

name	pdpco	pdpseq	gvkey	begyr	endyr
CELANESE CORP-OLD	162254	1	2827	1950	1985
HOECHST CELANESE CORP	162254	2	13934	1987	1996
CELANESE AG	162254	3	125434	1998	2002
CELANESE CORP	162254	4	162254	2003	2005

This represents a single corporate entity for our purposes. For the year 1992, we would use data from GVKEY = 13934.

3 Dynamic Assignee Match File

For each assignee, DYNASS contains up to five corporate matches:

```
Contains data from dynass.dta
  obs:        13,474                          Dynamic match of pdpass to
                                                Compustat gvkey/pdpco
 vars:            22                          7 Apr 2008 08:39
 size:       983,602 (98.5% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
pdpass          long   %12.0g                 New PDPASS - a few entities
                                                split
pdpco1          long   %12.0g                 Source id
source          str5   %9s
begyr1          int    %9.0g                  1 begyr
gvkey1          long   %9.0g                  1 gvkey
endyr1          int    %9.0g                  1 endyr
pdpco2          long   %9.0g                  2 pdpco
begyr2          int    %9.0g                  2 begyr
gvkey2          long   %9.0g                  2 gvkey
endyr2          int    %9.0g                  2 endyr
pdpco3          long   %9.0g                  3 pdpco
begyr3          int    %9.0g                  3 begyr
gvkey3          long   %9.0g                  3 gvkey
endyr3          int    %9.0g                  3 endyr
pdpco4          long   %9.0g                  4 pdpco
begyr4          int    %9.0g                  4 begyr
gvkey4          long   %9.0g                  4 gvkey
endyr4          int    %9.0g                  4 endyr
pdpco5          long   %9.0g                  5 pdpco
begyr5          int    %9.0g                  5 begyr
gvkey5          long   %9.0g                  5 gvkey
endyr5          int    %9.0g                  5 endyr
-------------------------------------------------------------------------------
Sorted by:  pdpass
```

The variables are:

PDPASS	The unique assignee number
SOURCE	The procedure used to make the match (for diagnostic purposes)
PDPCOi	The *i*th PDPCO matched to this assignee
GVKEYi	The *i*th GVKEY matched to this assignee
BEGYRi	The first year of the *i*th match
ENDYRi	The last year of the *i*th match

The multiple matches include both corporate reorganizations, such as Celanese, and also corporate mergers and acquisitions. For example, here is one record:

pdpco1	pdpass	source	begyr1	gvkey1	endyr1	pdpco2	begyr2	gvkey2	endyr2	pdpco3	begyr3	gvkey3	endyr3
3955	10949879	NBER	1966	3955	1997	3282	1998	3282	2001	5606	2002	5606	2005

(the fourth and fifth matches are missing).

The assignee is “DIGITAL EQUIPMENT CORPORATION”

PDPCO1 is “DIGITAL EQUIPMENT” in the Compustat file.

PDPCO2 is “COMPAQ COMPUTER CORP.” in Compustat.

PDPCO3 is “HEWLETT-PACKARD CO.” in Compustat.

Public data was available for Digital Equipment beginning in 1966. In 1998 Digital was a subsidiary of Compaq. In 2002, Compaq was a subsidiary of Hewlett-Packard.

4 Assignee file

```
Contains data from assignee.dta
  obs:       322,783
 vars:             5                          5 Apr 2008 09:38
 size:    53,259,195 (20.6% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
cod             byte   %34.0g      codlbl     Expanded assignment code for
                                                PDP data
cod_fix         int    %77.0g      fixlbl     reason for recoded assignee type
pdpass          long   %12.0g                 New PDPASS
standard_name   str150 %150s                  standardized assignee name
uspto_assignee  long   %12.0g                 PATSIC assignee number
-------------------------------------------------------------------------------
Sorted by:  standard_name
```

This file has a unique record for each name in standardized form used for each assignee. Note that there are multiple names hence multiple records for each of the different names the might refer to the same assignee (for example, “IBM” and “INTERNATIONAL BUSINESS MACHINES”).

The variables are:

COD	Assignee type
COD_FIX	Reason COD was changed (if changed) from original PTO type
PDPASS	Unique assignee number (may have multiple standard_names)
STANDARD_NAME	Standardized assignee name
USPTO_ASSIGNEE	Assignee number from PATSIC 04. Not all assignees have non-missing value. NOTE THAT THIS FIELD IS PROVIDED FOR BACKWARDS COMPATIBILITY ONLY. This field is missing for many assignees. Use PDPASS for a unique identifier of assignees.

The COD numbers are:

COD	Type	Number records
2	US corporation	134,889
3	Foreign corp, incl. state-owned	131,302
4	US individual	7,269
5	Foreign individual	37,560
6	US government	511
7	Foreign government	2,292
8	US local government	58
9	US state government	62
10	US university	1,830
11	Foreign university	1,887
12	US institute	1,188
13	Foreign institute	3,356

14	US hospital/med inst	405
15	Foreign hospital/med inst	174
	TOTAL	322,783

5 Patent-Assignee file

```
Contains data from patassg.dta
  obs:     3,032,482
 vars:             8                          14 Feb 2008 10:35
 size:   288,085,790 (45.1% of memory free)
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
sta             str2   %2s                    assg/state
cnt             str3   %3s                    assg/country
assgnum         byte   %8.0g                  assg/assignee seq. number (imc)
cty             str72  %72s                   assg/city
pdpass          long   %12.0g                 New PDPASS
ptype           str1   %9s                    patent type
patnum          long   %12.0g                 patent number
------------------------------------------------------------------------------
Sorted by:  ptype  patnum
```

This file contains a record for each patent:assignee pair for all assigned patents from 1976-2006. Note that there may be multiple assignees per patent (enumerated by ASSGNUM). The variables are:

sta	Assignee state
cnt	Assignee country
assgnum	Assignee sequence number for this patent (multiple assignees)
cty	Assignee city
pdpass	Unique assignee number
ptype	Patent type (see below)
patnum	Patent number (stripped of leading alphabetic type designator)

The patent types and counts are:

ptype	Type	Number
0	Utility	2,812,428
D	Design	198,114
H	Statutory Invention Registration	1,812
P	Plant	10,625
R	Reissue	9,281
T	Defensive publication	222
Total		3,032,482

6 Sample Code

Suppose one has two files:

WORK which contains financial data for each GVKEY-YEAR, sorted by GVKEY and YEAR

NPAT which contains the number of patents, NPAT, for each PDPASS-YEAR sorted by PDPASS and YEAR.

This code adds counts of NPAT to work:

```
use npat, clear
* merge dynamic assignee data into pdpass-npat file
merge pdpass using dynass
tab _m
keep if _m==3

* now find the appropriate gvkey to assign the patents
gen gvkey=.
forvalue i=1/5 {
      replace gvkey = gvkey`i' if gvkey`i'~=. & year>=begyr`i' &
year<=endyr`i'
}

keep if gvkey~=.
keep gvkey year npat

* sum over multiple assignees to get patents for each company
sort gvkey year
collapse (sum) npat,by(gvkey year)

* merge in the work data
merge gvkey using work
tab _m
drop if _m==1
drop _m
```

We are not quite done yet, however, because we know that some of the firms in WORK have zero patents (as opposed to NPAT = missing). This does that

```
* merge in match variable
sort gvkey year
merge gvkey using pdpcohdr
tab _m
drop if _m==2
drop cusip name firstyr lastyr pdpco pdpseq begyr endyr _m

* create match flag variable
gen mtchflg= match~=.
replace npat = 0 if mtchflg & npat==.
drop match
```

If one wants to calculate patent stocks, as opposed to simple patent counts, using a perpetual inventory method, one has to first calculate patent stocks for each pdpass for each year and then merge these data into the WORK file using similar code. Note that it will not work correctly to build patent stocks from the npat data because assignees acquire patents during years not necessarily captured in the WORK

file.

7 Quality of the Match

[to come]

8 Data release alpha3, April 2008

For details on the development of the data, please consult the programmer documentation.

In summary, the main sources of initial data are:

Compustat firm data	WRDS data from 1950-2006 captured in CSHDR06.DTA
Patent assignee data	CONAME04.DTA from the USPTO PATSIC4 product and the ASSG file from the USPTO through 2006
Initial subsidiary information	COMPUSTAT.TXT from NBER file based on 1989 *Who Owns Whom?*
Merger and acquisition data	SDC data from 1975-2006

These data are combined with manual and computer-assisted generated data to generate the initial build. This initial build was then modified using the update procedure described in the programmer documentation. The sources of the records in DYNASS are:

Source code	Number records	Description of source
FRAT	181	Updates to the original NBER file based on data compiled by Annette Fratantaro (Federal Reserve Bank of Philadelphia)
MTCH1	4,398	Manual matches made March 2007
NBER	3,765	Original matches in NBER 1989 file
PASS1	3,846	Computer generated matches of standardized names
UPDT2	497	Manual matches in update of June 2007
m2006	787	Matches based on 2005-6 data, April 2008
Total	13,474	