Patent Data Project

patn data description

The file pdp76_06 has been revised as of August 2010. The following are important changes:

The variable *allcites* is NOT corrected for citation truncation (in the previous version it was) but the correction factor *hjtwt* is included to allow the user to correct.
The variable *allnscites* has been removed, as it is only really useful if you have a specific dataset of US companies, and was inaccurate otherwise.
New variables have been added:

nclaims - number of claims in the patent (from the website, probably not corrected for any re-exams or disclaimers)
status - whether the patent is missing (M) or withdrawn (W). There are very few of these; they often will not have any other data on the file.
term_extension - the number of days that the patent term has been extended due to USPTO delays.

The *cat, subcat, nclass*, and *subclass* variables now pertain to the current classification (CCL), not the original classification (OCL). The variables corresponding to the original classification are labelled *cat_ocl*, etc. This change was made because the CCL is more complete and probably more accurate about the current state of the system.

Finally, this file includes multiple records for patents with multiple assignees. There are 63,266 such patents, with up to 13 assignees. This is about 2 per cent of the total number, and increases the number of observations in the file from 3,210,361 to 3,279,509. For patent analysis, you will wish to remove the duplicate observations, but to count patents held by an entity, you may wish to keep all the entries. Other than *assgnum* (which is an older and inaccurate sequence number), *pdpass, state,* and *country*, all data for a given patent number are identical.

Patent classification:

The USPTO changes its classification system from time to time to accommodate the growth in new technologies, adding classes, and occasionally deleting old classes if they become too full (creating a whole new set of classes to replace them).

The variables designated "OCL" are based on the original classification at the time the patent was examined and issued (the field of search shown on the USPTO website). Therefore the classificaiton system used will vary across the patents in the file according to their vintage.

The variables designated "CCL" are based on the USPTO classification system as of 2008. This means that all the patents on this file will have a consistent classification applied if you use the ccl variables. The category and subcategory assignments are listed in this spreadsheet. Note that the categories do not fully correspond to early technology classes (such as 2006 or nclass) because, of course, the USPTO continually revises the technology classes.

IPC codes are assigned via a concordance from the USPTO codes. The USPC to IPC Concordance is based on the International Patent Classification Eighth Edition (please see note at the bottom of this page).

Correcting for citation truncation:

HJT refers to Hall, Bronwyn, Adam Jaffe and Manuel Trajtenberg, "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," NBER Working Paper 8498. This paper describes the citation variables, their computation and weights. The variable hjtwt is a multipler that can be applied to the number of citations from US patents through 2006 received by the patent (allcites), in order to correct for the truncation of post-2006 cites using the methodology described in the HJT data description. HJT estimated a 6 field specific obsolescence-diffusion model with year and lag dummies and used the estimated model to predict a grossing up factor for the cites based on the patent's grant year and technology category. As discussed in the HJT paper, this variable is not very accurate for the last three years of the sample (2004-20066), as three years is too short a time to get a good measure of actual cites.

The formats of the new utility patent data files are:

1. pat76_06_assg.dta (and ASCII equivalent)

This file has one record for each *assignment* of each utility patent. Patents that are assigned to more than one party have multiple records. This file lists only the first technology class.

The data description of this file is:

```
  obs:     3,279,509                          Patents granted
through 2006

 vars:            32                          4 Aug 2010 14:25

 size:   377,143,535 (28.1% of memory free)

----------------------------------------------------------------------

              storage  display    value

variable name   type   format     label      variable label

----------------------------------------------------------------------

allcites        int    %9.0g                 Cites 1976-2006 (not
adj for

                                               truncation)

appyear         int    %8.0g                 Year patent applied
for

asscode         byte   %8.0g                 Original assignee code
(1-7)

assgnum         byte   %8.0g                 assg/assignee seq.
number (imc)
```

```
cat             byte   %8.0g                  HJT tech category
(1-6) for CCL

cat_ocl         byte   %8.0g                  HJT tech category
(1-6) for OCL

cclass          str11  %11s                   Primary current US

                                                class/subclass
(Alpha)

country         str2   %9s                    assg/country

ddate           float  %d                     patn/disclaimer date

gday            byte   %8.0g                  Day patent granted

gmonth          byte   %8.0g                  Month patent granted

gyear           int    %8.0g                  Year patent granted

hjtwt           float  %9.0g                  Citation truncation
weight as

                                                of 2006

icl             str18  %18s                   clas/international

                                                classification

icl_class       str4   %9s                    Main 4-char IPC

icl_maingroup   float  %9.0g                  Main group within
4char IPC

iclnum          byte   %8.0g                  clas/icl seq. number
(imc)

nclaims         int    %9.0g                  patn/number of claims

nclass          int    %9.0g                  US 3-digit current

                                                classification (CCL)

nclass_ocl      int    %9.0g                  US 3-digit original

                                                classification (OCL)

patent          long   %12.0g                 Patent number (7-
digit)

pdpass          long   %12.0g                 Unique assignee number

state           str2   %2s                    assg/state

status          str1   %1s                    auth/status: m
missing, w

                                                withdrawn

subcat          byte   %8.0g                  HJT tech subcategory
(11-69)

                                                for CCL

subcat_ocl      byte   %8.0g                  HJT tech subcategory
(11-69)
```

```
                                                for OCL

subclass        float  %9.0g                  Subclass for US
current class

                                                (CCL)

subclass1       str9   %9s                    Subclass for US
current class

                                                (CCL) - Alpha

subclass1_ocl   str9   %9s                    Subclass for US
original class

                                                (OCL) - Alpha

subclass_ocl    float  %9.0g                  Subclass for US
original class

                                                (OCL)

term_extension  int    %9.0g                  patn/extension of
patent term

                                                in days under 35 usc
154(b)

uspto_assignee  long   %12.0g                 Original assignee
number

----------------------------------------------------------------------

Sorted by:  patent
```

2. pat76_06_ipc.dta

This file has one record for each IPC class for each patent. The data description is:

```
  obs:     4,857,833                         Patents granted
through 2006,

                                                including IPCs

 vars:            14                          4 Aug 2010 14:27

 size:   272,038,648 (48.1% of memory free)

----------------------------------------------------------------------

              storage  display     value

variable name   type   format      label      variable label

----------------------------------------------------------------------

appyear         int    %8.0g                  Year patent applied
for

cat             byte   %8.0g                  HJT tech category
(1-6) for CCL

gyear           int    %8.0g                  Year patent granted

icl             str18  %18s                   clas/international
```

```
                                                  classification

   icl_class       str4   %9s                   Main 4-char IPC

   icl_maingroup   float  %9.0g                 Main group within
   4char IPC

   iclnum          byte   %8.0g                 clas/icl seq. number
   (imc)

   nclass          int    %9.0g                 US 3-digit current

                                                  classification (CCL)

   numipc          byte   %9.0g                 Number of
   international patent

                                                  classes

   patent          long   %12.0g                Patent number (7-
   digit)

   pdpass          long   %12.0g                Unique assignee number

   subcat          byte   %8.0g                 HJT tech subcategory
   (11-69)

                                                  for CCL

   subclass        float  %9.0g                 Subclass for US
   current class

                                                  (CCL)

   uspto_assignee  long   %12.0g                Original assignee
   number

   ----------------------------------------------------------------------

   Sorted by:  patent  pdpass  iclnum
```

classification_06.xls (187k)   Jim Bessen, …   v.1

Comments

You do not have permission to add comments.