

Patent Data File

Peter Thompson and Melanie Fox Kean (2005): "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment" *American Economic Review*.

This tab-delimited file contains the data used in the paper. It consists of all patents with an institutional assignee granted during the first week of January, 2003, and all patents they cite granted after January 1, 1976. The citing patents appear multiple times, once for each citation they make. The columns of the files are:

Citing number	Number of the citing patent.
Cited number	Number of the cited patent.
Added by examiner	Equals 1 if the examiner added the citation, 0 otherwise
Cited noninstitutional	Equals 1 if the assignee of the cited patent is a company, university, or government entity, 0 if the assignee is a private individual or no assignee other than the inventor is given.
Self-citation	Equals 1 if the citation is to a patent with the same assignee.
*Citing file date	Filing date of citing patent.
*Citing issue date	Issue date of citing patent.
*Citing assignee name	In the case of multiple assignees, the first-named assignee was used. Where possible, the file name was converted to the preferred format given in the NBER patent files. This variable needs further manual cleaning before it can be used reliably.
*Citing assignee city	The city of the assignee as given in the patent. I did not use this variable in the paper, so it has not been cleaned up. Some place names are missing, and foreign place names (especially Taiwanese) have all sorts of variants of spelling. This variable needs further manual cleaning before it can be used reliably.
*Citing assignee state	The state of the assignee as given in the patent. I did not use this variable in the paper, so it has not been cleaned up. In particular, the state code often confounds countries and US states. For example, DE is Delaware and Germany, CA is California and Canada. It is tedious to clean these up by hand, but it can be done. While the html version of the patent on the USPTO web site does not clarify, the image file does. Until 2001, for example, the patent image shows "Calif." and "Canada" while the html file uses "CA" for both. After January 2001, the image gives "CA (US)" and "CA" for the state and county respectively. This variable needs further manual cleaning before it can be used reliably.
*Citing first inventor city	The city of residence of the first-named inventor. Please note that this variable has not been cleaned up for foreign locations. Foreign place names (especially Taiwanese) have all sorts of variants of spelling, and often other parts of a mailing address are still included.
*Citing first inventor state	The state of residence of the first-named inventor. Noted as "Foreign" if non-US.
*Citing first inventor county	The county of residence of the first-named inventor. Noted as "Foreign" if non-US. See notes to correlation file above.
*Citing first inventor msa	The MSA of residence of the first-named inventor. Noted as "Foreign" if non-US. See notes to correlation file above.

- *Citing first inventor cmsa The CMSA of residence of the first-named inventor. Noted as "Foreign" if non-US. See notes to correlation file above.
 - *Citing first inventor country The country of residence of the first-named inventor.
 - *Citing examiner The name of the primary examiner.
 - *Citing US class The primary US classification code.
 - *Citing Int class The International classification code. Where more than one code is given in the patent, the first-listed is given here.
- * All fields with an asterisk are then repeated in the same order for the cited patent.

Correlation File

This tab-delimited file maps place names to metropolitan statistical areas (MSAs) and consolidated metropolitan statistical areas (CMSAs) as defined by the 1990 tables of the Bureau of the Census. The file is built upon the correlation tables supplied by [Office of Social and Economic Data Analysis \(OSED\)](#) of the University of Missouri. I have added to these data about 1,000 place names that do not appear in the OSEDA files, but that do appear in the patent data set. To locate these place names, I used Yahoo maps to find the the nearest place name that does appear in the OSEDA files, and copied its county and MSA. In a small number of cases, this will assign the incorrect county (because the nearest place -- almost always within 5 miles -- was just the other side of a county line). The MSA will always be correctly recorded, however. I have also added an entry "CMSA/Phantom", motivated by the geographic classifications scheme of Jaffe Trajtenberg and Henderson's original work on citations. This lists the CMSA, if the place belongs in one, or "Phantom-XX" if the place does not, where "XX" is the two-letter postal code for the state. Finally, I have added many spelling variants as they appeared in my patent data set. I have found that using these tables in a relational database will automatically generate county, MSA or CMSA locations for about 95 percent of the place names in a large patent dataset. Many of the remainder will turn out to be misspellings of places in the file and must be cleaned by hand. You can expect that less than 1 percent of your observations will be place names that do not appear in this file. The columns of the file are:

State -- two-letter postal code

Place -- placename (cty, town, neighborhood)

County -- note the caveat in the text above.

MSA -- there are 300+ MSA's defined. Rural locations not assigned to an MSA by the Census Bureau are recorded as "Non-metro"

CMSA -- there are 17 CMSA's, each of which consists of multiple MSAs. Many MSAs are not part of a CMSA. Places not assigned to a CMSA are codes as "Phantom-XX", where XX is the two-letter state code.