Merge Sum Stats

Sylvia Klosin

Merge Summary Statistics

Filter

In this case we filter the patent and infutor datasets to include only

- People whose last name starts with "DI"
- Observations more recent then 2000 in the patent data
- People who live in the United States

Merge

We merge on last name, state, city and the first three letters of the first name. We have 3706 unique inventors that we are trying to find matches for. The corresponding infutor set has 3133015 unique person ids.

1-1 matches

We find that doing this merge we get 785 inventors that do not get any matches in the infutor data.

(We were able to lower this number a few different ways. 1) we expanded the infutor time period 2) we cleaned last names by removing things like "JR" and "III".)

There are 1947 inventors that get exactly one match.

For these 1-1 matches, we get 0.927784 have the same clean first name. For those 7% where there is not an exact match, looking at the names show that they seem to be cases where nicknames are used e.g. Fredrick vs Fred.

> 1 matches

Full First Name

As a next step we bring in full first name information. Adding in the first names gives us 172 more matches. We have a total of 2119 1-1 matches.

Next time we bring in first letter of middle name

Checking dates gives us a total of 2335 1-1 matches. Now we have 500 greater than 1 matches.

Now we check bad dates

We check if date of application and date of infutor are within plus or minus 180 days.

Checking dates gives us a total of 2480 1-1 matches. Now we have 355 greater than 1 matches.

Checking Matches 1-1 matches

Looking at dates

Seeing for the matches, is there at least one "true" year match. That at least one of the year/addresses in the patent data matches a year/address in the infutor data. We find that this is NOT the case for 514 inventors. The flexibility given is 1 year.

	V1
2000	39
2001	68
2002	68
2003	39
2004	86
2005	40
2006	76
2007	34
2008	24
2009	12
2010	7

Second Round Merge to get the people who did not match at all

Looking at the matches when infutor only has first letter of first name gives us only 5% more matches.