Patent re-classification and likelihood of IPR

Matthew Baker

12/11/2019

Part 1 - Introduction

This project will explore the outcomes of certain patents issued by the United States Patent Office (USPTO).

The first group of patents is one which tracks variables of all patents including re-classification and number of citations.

At the time of issuance, all patents are assigned a classification code that describes the subject matter of the patent. Afterwards researches can find patents by searching within pertinent classification codes. This system is called the US Patent Classification (USPC). The USPC is periodically revised to account for the evolution of technologies. As innovation in certain technologies slows to a halt (e.g. Muzzleloader firearms, Credit Card imprinters, Typewriters, Fax Machines), the classification is updated and certain patents must be re-classified.

The second group of patents is one which tracks legally challenged patents.

After a patent is issued its validity can be legally challenged. Challenges can be filed either in federal court or at the USPTO itself.

My hypothesis is that higher visibility will lead to higher rate of challenge. The independent variable, both treated categorically will bereclassification (yes/no). The outcome variable will be if a patent was later challenged (yes/no).

Part 2 - Data

```r
#For ease of analysis the data is downloaded and in csv format (one of the files is too large to host o

# The first dataset contains five variables: 1) patent: patent number 2) gyear: grant year
#3) nclass_ocl: original (at birth) main class 4) class: current main class 5)
#allcites: number of citations received between birth and June 2015.


#clean data, for the purposes of analysis the patents starting with D are ignored, these are
#design patents and will not show up in the compared dataset
patent_c$nclass_ocl <- trimws(patent_c$nclass_ocl, which = c("both"))
patent_c$nclass <- trimws(patent_c$nclass, which = c("both"))

#compare data and create new comparison qualitative variable
patent_c$changed<-as.numeric(as.numeric(patent_c$nclass) != as.numeric(patent_c$nclass_ocl))
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```r
head(patent_c)
```

```
##    patent gyear nclass_ocl nclass allcites changed
## 1 RE28671  1976        009    441        0       1
## 2 RE28672  1976        016    016        9       0
## 3 RE28673  1976        053    053        0       0
## 4 RE28674  1976        128    604       23       1
## 5 RE28675  1976        180    180        1       0
## 6 3930271  1976        002    002       14       0
```

```
#second dataset challenged patents list
patent_ptab$challenged<-as.numeric(1)
head(patent_ptab)
```

```
##     TrialNumber ApplicationNumber FilingDate PatentNumber challenged
## 1 CBM2014-00042          09703562 2013-12-05      7499872          1
## 2 IPR2016-00623          14034206 2016-02-12      8873500          1
## 3 IPR2016-00644          09486648 2016-02-23      7010572          1
## 4 IPR2017-01346          12046099 2017-05-05      8161344          1
## 5 IPR2017-01429          10685266 2017-05-12      7214506          1
## 6 IPR2017-01511          10984366 2017-05-31      7242028          1
```

#merge datasets and wrangle into categorical variables

```
alldata<-merge(patent_c, patent_ptab, by.x = "patent", by.y="PatentNumber", all=TRUE)

tabledata<-alldata[c("patent", "changed", "challenged", "allcites")]
tabledata %>% mutate_if(is.factor, as.character) %>%
  filter(!grepl('^D', patent)) %>%  filter(!grepl('^T', patent)) %>% replace(is.na(.), 0) %>% mutate(al
```

```
## Warning: Trying to compute distinct() for variables not found in the data:
## - `tabledata`
## This is an error, but only a warning is raised for compatibility reasons.
## The operation will return the input unchanged.
```

```
head(tabledata)
```

```
##    patent changed challenged allcites
## 1 3930271       0          0        1
## 2 3930272       0          0        1
## 3 3930273       0          0        1
## 4 3930274       1          0        1
## 5 3930275       0          0        1
## 6 3930276       0          0        1
```

The data was collected as shown above, by merging two separate datasets and using the patent number as a key value. Then the independent variable under study was made into a binary categorical variable. As shown below there ends up being 5128366 observations of 4 variables. The number of re-classified patents in the data is 404146 (this is the categorical explanatory variable). However the challenge subset is much smaller and there are 7605 challenged patents (this is the categorical response variable). This will be an observational study. Links for the data are provided in the References section. Because two categorical variables are being analyzed I will use a chi-square analysis. This method could be generalized to other characteristics of patents which are binary and categorical (such as Foreign vsDomestic inventor), or with slight modification it could be applied to categorical variables with more than one possible value (such as inventor country of origin). As for the scope of inference, it will be interesting to see how big the effect of the explanatory variable is. Chi Square test can certainly show correlation but two establish casuality more than two matched datasets are required, this will have to wait.

The public benefit of patent validity is important because it is a system built on an quid pro quo that requires much investment from the public in exchange for a period of monopoly.

Part 3 - Exploratory data analysis

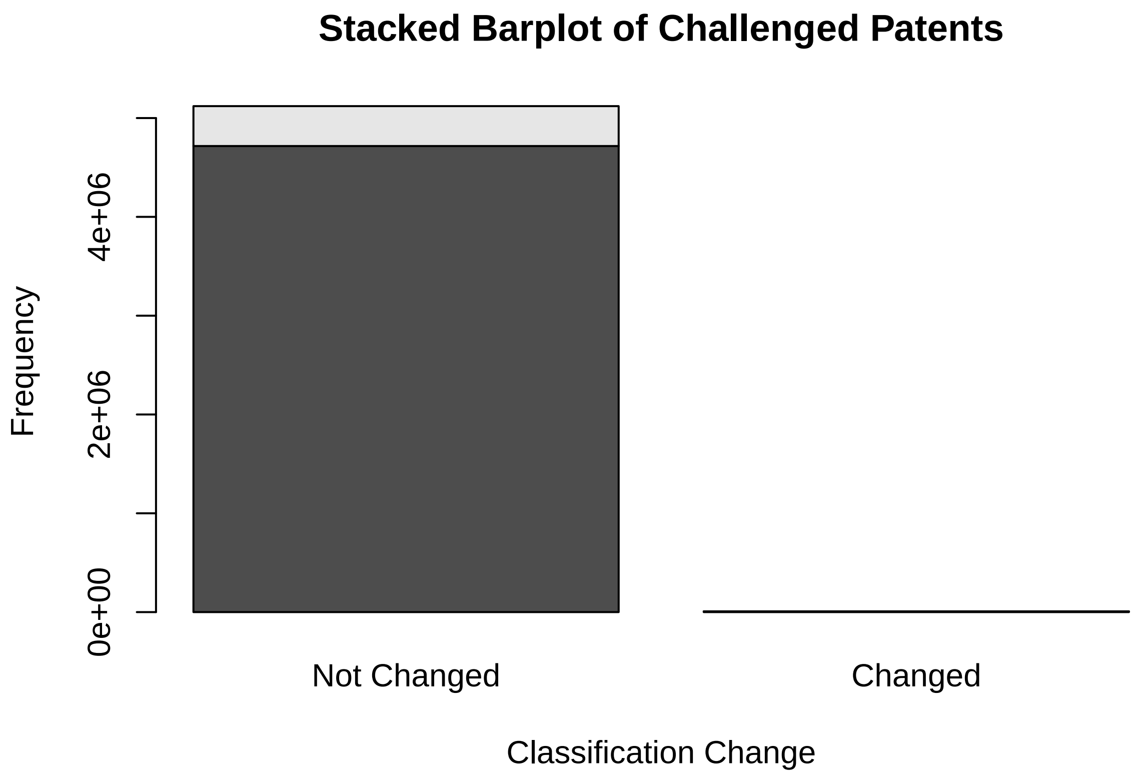
##Graphs

The barplot highlights that challenge is a very rare event.

```
#focusing on classification change and challenge outcome
mdata<-tabledata[c("changed", "challenged")]
table(mdata)
```

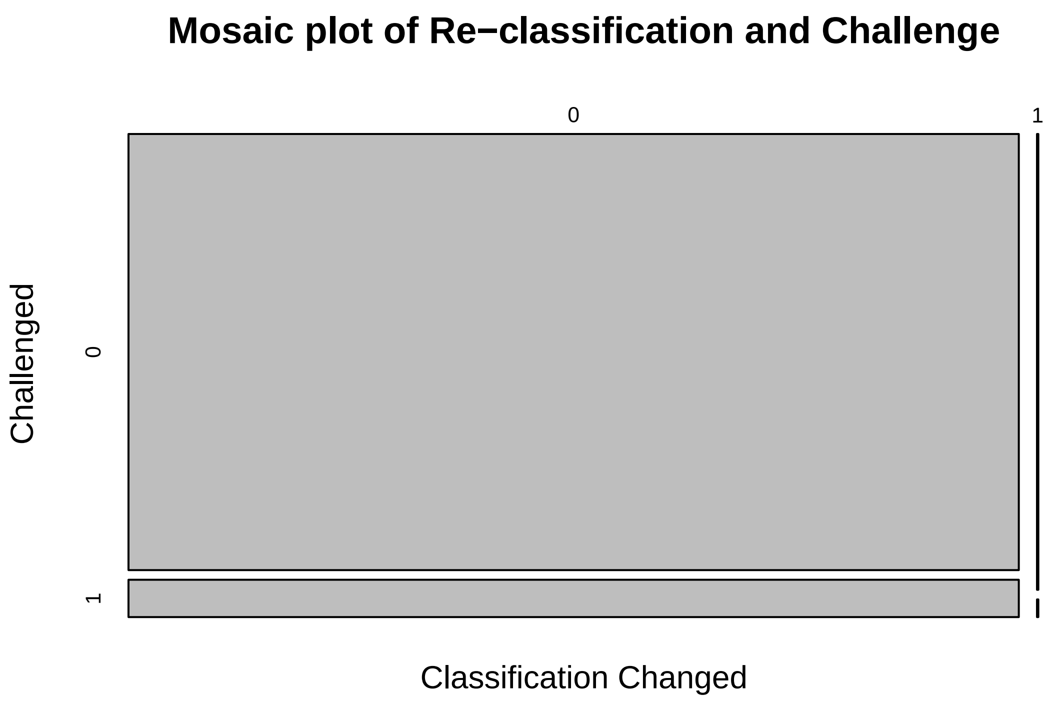
```
##        challenged
## changed       0       1
##       0 4716973    7281
##       1  403830     282
```

```
names<-c("Not Changed", "Changed")
barplot(table(mdata), main = "Stacked Barplot of Challenged Patents", xlab ="Classification Change", yl
```



The mosaic plot shows the difference in the proportions, detailed below.

```
mosaicplot(table(mdata$challenged,mdata$changed), main = "Mosaic plot of Re-classification and Challeng
```



##Descriptive statistics
Contingency Table, shows how rare a challenge is

```
mdata.tab<-table(mdata$changed,mdata$challenged)
mdata.tab
```

```
##
##           0       1
##   0 4716973    7281
##   1  403830     282
```

```
282/7281
```

```
## [1] 0.03873094
```

```
403830/4716973
```

```
## [1] 0.08561211
```

##proportional table of all cells
again shows how rare a challenge is. this table shows that reclassification is not common however it occurs in around 8.56% of all patents

```
t1<-round(prop.table(mdata.tab),6)
t1
```

```
##
##            0        1
##   0 0.919781 0.001420
##   1 0.078744 0.000055
```

##proportional table of rows

```
round(prop.table(mdata.tab,1),6)
```

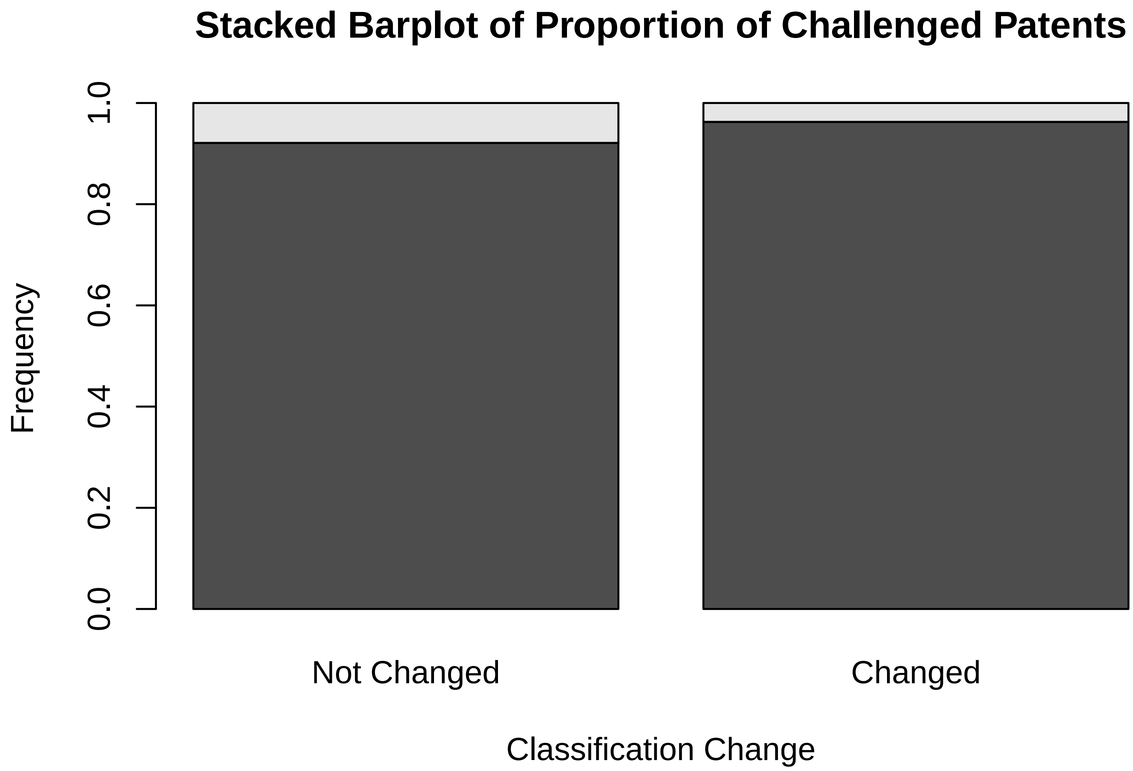
```
##
##            0        1
##   0 0.998459 0.001541
##   1 0.999302 0.000698
```

##proportional table of cols this is contrary to what I expected, which was that re-classification would increase the rate of challenge. however, this shows the opposite.

```
t2<-round(prop.table(mdata.tab,2),6)
t2
```

```
##
##            0        1
##   0 0.921139 0.962713
##   1 0.078861 0.037287
```

```
barplot((as.matrix((t2))), main = "Stacked Barplot of Proportion of Challenged Patents", xlab ="Classif
```



Part 4 - Inference

Does the distribution of challenged patents differs when re-classified? The sample observations are certainly independent as challengers are not related. Frequency is high enough and data has been constrained. H0: the ratio of reclassified patents in challenged and non-challenged sets is the same H1: the ratio is different

```
chisq.test(mdata.tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  mdata.tab
## X-squared = 179.24, df = 1, p-value < 2.2e-16
```

Part 5 - Conclusion

Chi Sq is very high and p value is very low, and statistically significant. The variations in the re-classification rates of patents are bigger than expected when comparing the challenged and non-challenged populations, but with the opposite effect that I hypothesized. Reject the H0.

The challenge rate is smaller than initially expected from a cursory review of the data, so other similar datasets could prove helpful (such as federal court challenges instead of USPTO challenges).

This type of study will bring better understanding if the outcome of the challenge itself is categorized (such as patent held valid, patent nullified, or parties settlement agreement entered.)

References

#Paper that inspired the project https://link.springer.com/article/10.1007/s00191-018-0603-3

second data set (requires google sign-in) https://bigquery.cloud.google.com/table/patents-public-data:uspto_ptab.trials_201710?pli=1