An Analysis of PERM Labor Certification and Labor Condition Applications from the United States Department of Labor

Arunkumar Ranganathan
Brian Detweiler
Jacques Anthony

October 18, 2016

Abstract

Foreign born workers make up 17% of the United States workforce. In 2014, nearly one million foreign nationals became lawful permanent residents in the United States. Of those one million, 140,000 came through visas which are allocated to employment based residency. Where are these workers, and what do the demographics look like? How does each company's compensation measure up? Here, we use statistical analysis and business analytics to examine visa application data from the U.S. Department of Labor from 2008 to 2016. We intend to create an interactive data product that will make this publicly available information more accessible to the students who are entering the workforce, as well as to US citizens and permanent residents. This will empower them to competitively position themselves in the job market by making more informed decisions.

1 Introduction

The U.S. Department of Labor provides data for Labor Condition Applications and PERM Labor Certifications dating back to 2008. This data contains a wealth of job market information including prevailing wage, and the wages offered by particular companies to individuals with particular qualifications.

1.1 Document reproducibility

The entirety of this project is reproducible.

2 About the data

The Office of Foreign Labor Certification, under the Department of Labor provides data for PERM Labor Certification (LC) applications and Labor Condition Applications (LCA) via XLSX files. Data is available from 2008 onward. The iCERT system was implemented in 2009, so there are two files for 2009. Each file is structured similar to the others, but there are differences which must be addressed.

2.1 H-1B Data preparation

The H-1B data is about 75% larger than the PERM data, and spreadsheet programs do not handle these well, so the first task was to export these to CSV formatted files so that they could be handled with better tools. When exporting, they were also given more uniform file names. Once in CSV, we needed to identify common columns across all spreadsheets. The difficulty here, is that the columns do not have the same names across spreadsheets, even though they may be holding the same data.

Using the UNIX tool `head -n 10 *.csv > headers.txt`, we took the first ten rows of each file and put them into a separate file. Each of the CSVs first ten rows were then copied and pasted into another spreadsheet, and we undertook a manual effort to match columns of the same identity. We also discarded some excess information that we deemed to be unnecessary for our purposes.

It is also important to note that there was not always a match for the columns we had selected. For instance, we found some interesting information regarding the attorney used by the employer to file the H-1B application. This was only introduced in the 2015 and 2016 datasets, however, so prior years would have no data for this.

After determining the standard columns, we wrote an import script in R that made use of the `data.table::fread` function. This allowed us to not only quickly read in the file, but also select only the columns of interest, and rename them to the standard naming convention upon read.

Once the data was read into individual data frames, additional cleaning rules were applied. In some years, wage data contained invalid numeric characters such as dollar signs or a range of wages in a single column. To get around this, dollar signs were removed before converting to numeric, and ranges were split into a *from* and a *to* column. Ultimately, all wages were transformed into ranges. If there was no range for the wage, then the wage itself was used as the range.

Another normalization task was the wage unit. Some wages were represented as yearly salary, some as hourly, and others as monthly, weekly, and bi-weekly. There can be subtle differences in each type of pay, but these were normalized according to a yearly salary. Hourly wage was multiplied by 40 hours a week and 52 weeks a year. Monthly wage was multiplied by 12, weekly by 52, and bi-weekly by 26. This allows all wages to be treated roughly on the same scale.

2.2 Shortcomings

As mentioned in the previous section, because the data is not homogeneous, there are bound to be disparities. Missing data - columns which are not found across all spreadsheets - is the biggest issue. We can make assumptions when there is sparse data, but it would not be prudent to make assumptions where there is no data. For this reason, we fully disclose the absence data where necessary.

All of the data has been entered by humans at some point, so there are likely many human-generated errors. Some of these can be seen as outliers. Particularly in the 2008 and pre-iCERT 2009 data, the wage unit is most certainly incorrect in some spots. For example, some wages are listed at $500 per hour, but the intended unit may have been per week. It is not possible to fix this programmatically though, because there are, in fact, some jobs that pay $500 per hour (CEOs, for instance). This data must be dealt with in one of two ways. They can either be corrected by hand inspection of outliers, or outliers can be removed completely. This results in a slight loss of fidelity. Extremely high paying jobs, such as CEO or physician may not be displayed.

Another issue is the switch from U.S. Citizenship and Immigration Services Dictionary of Occupational Titles (DOT) codes in 2008 and pre-iCERT 2009 data, to the North American Industry Classification System (NAICS) codes. The DOT codes are three digits and fairly high-level, where as the NAICS codes are hierarchical, with the first two being the industry, and the specification of the job title narrowing with up to six digits. For this reason, it is difficult to get consistent job titles across years.

3 Methods

The numeric data provided by the H-1B and PERM datasets are mostly in the form of wages, both the wage that the employer is offering and the prevailing wage.[1] Also of interest, are the number of workers an applicant is filing for, and the implicit number of applications faceted by status and year.

3.1 Data product

The data is provided as an interactive `Shiny` application, that allows the user to filter wages by various criteria.

The plots consist of distributions and and heat maps of wages across the United States.

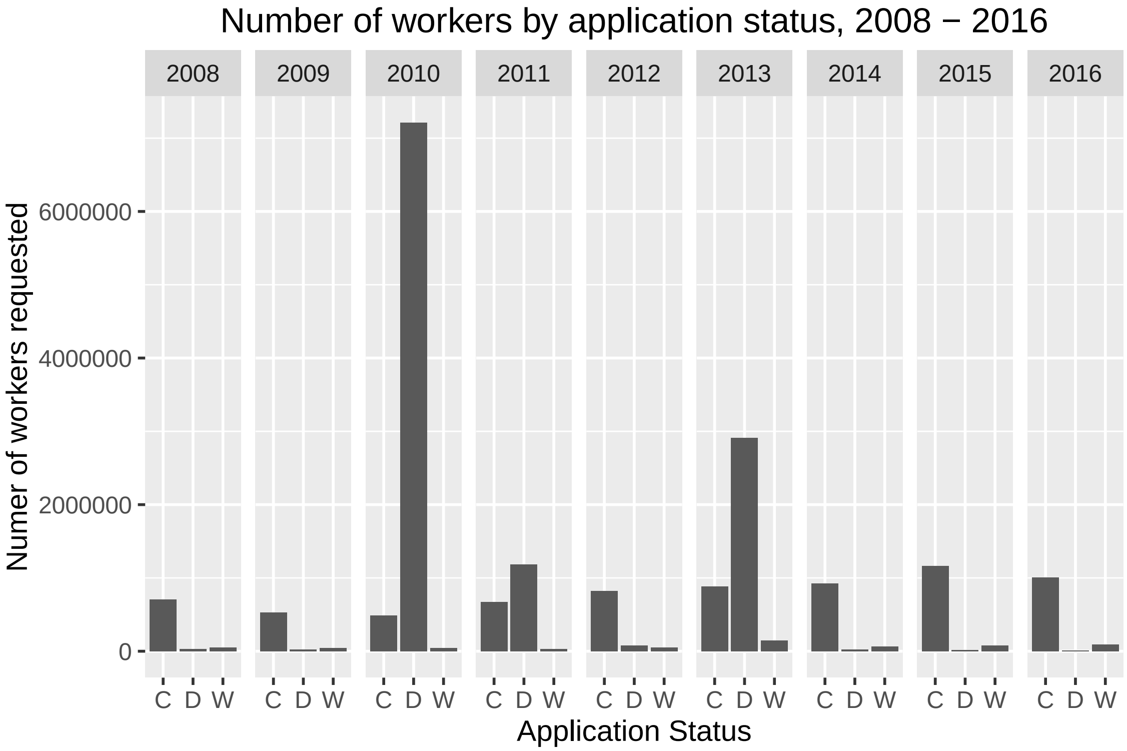
[1] Prevailing wage is defined as the hourly wage, usual benefits and overtime, paid to the majority of workers, laborers, and mechanics within a particular geographic area.

4 Results

4.1 Overview

4.2 Prevailing Wage

It would be interesting to see how many workers are getting approved for H-1B visas over the years. We can see this by plotting the number of workers within each visa status category (Certified, Denied, and Withdrawn).

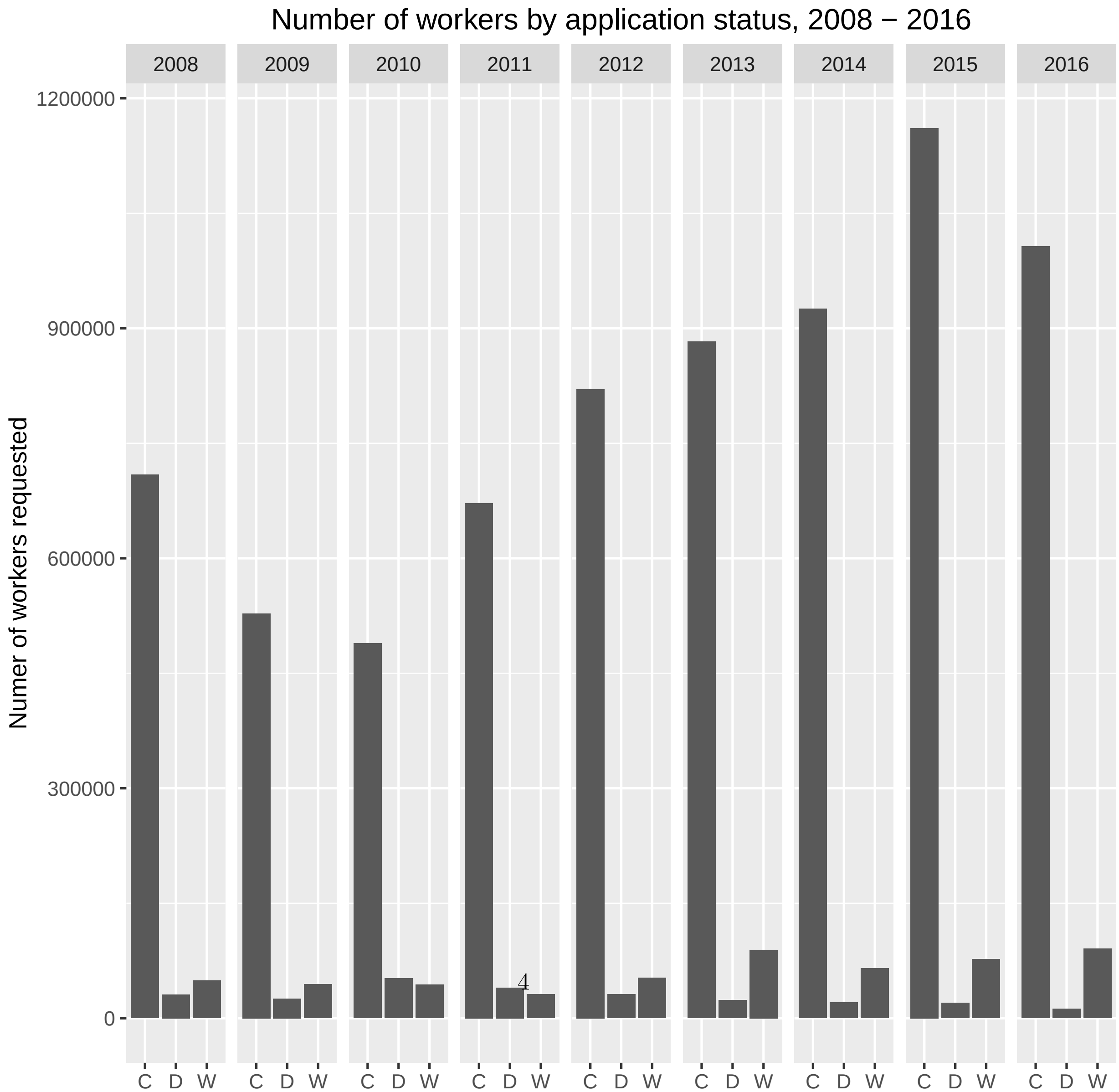


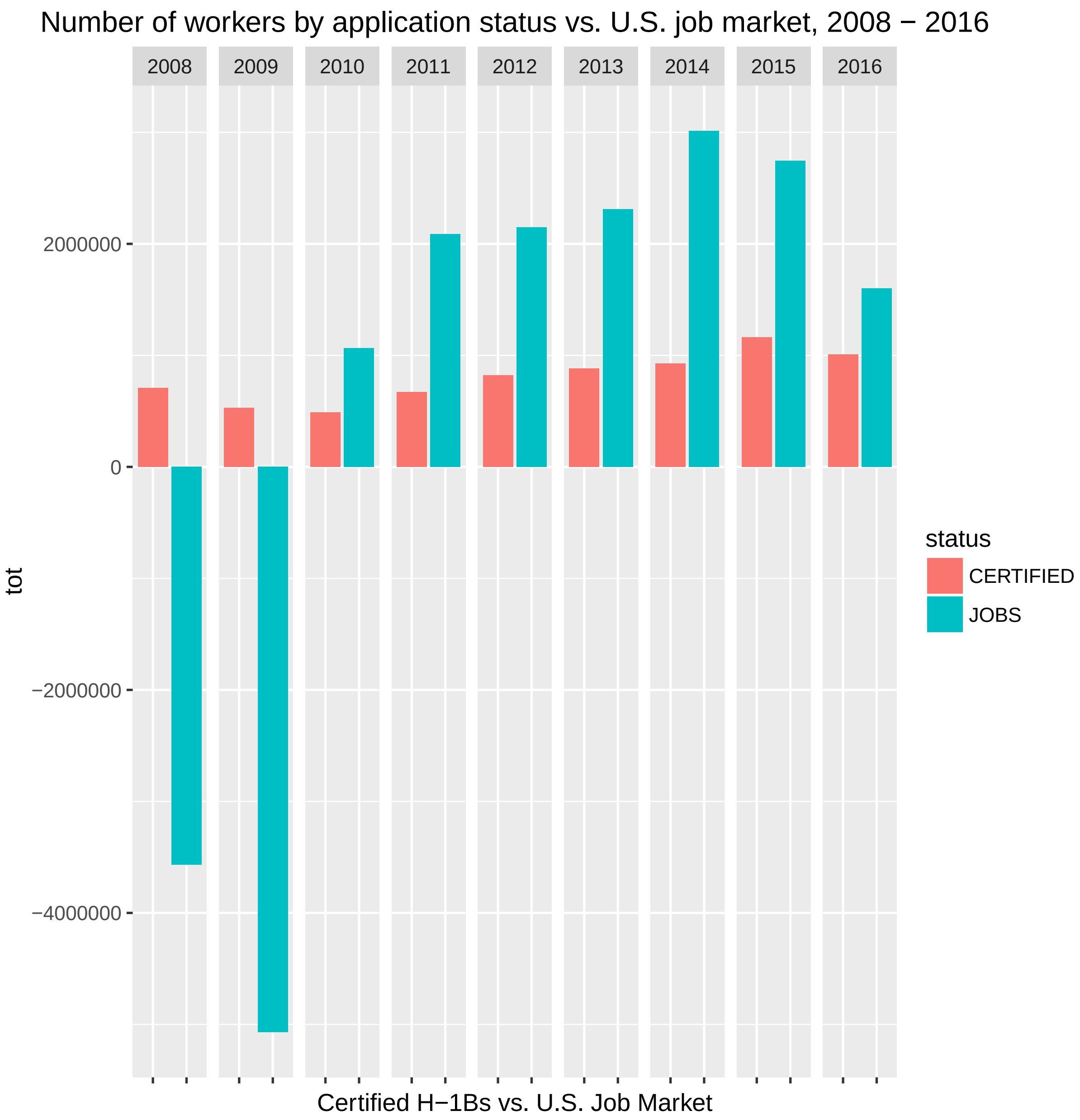
Clearly, we can see huge outliers in 2010 and 2013 that don't seem to fit the data. These could be explained as outliers. Upon investigation, it is safe to say that all H-1B applications requesting over 1,000 total workers are either denied or withdrawn.

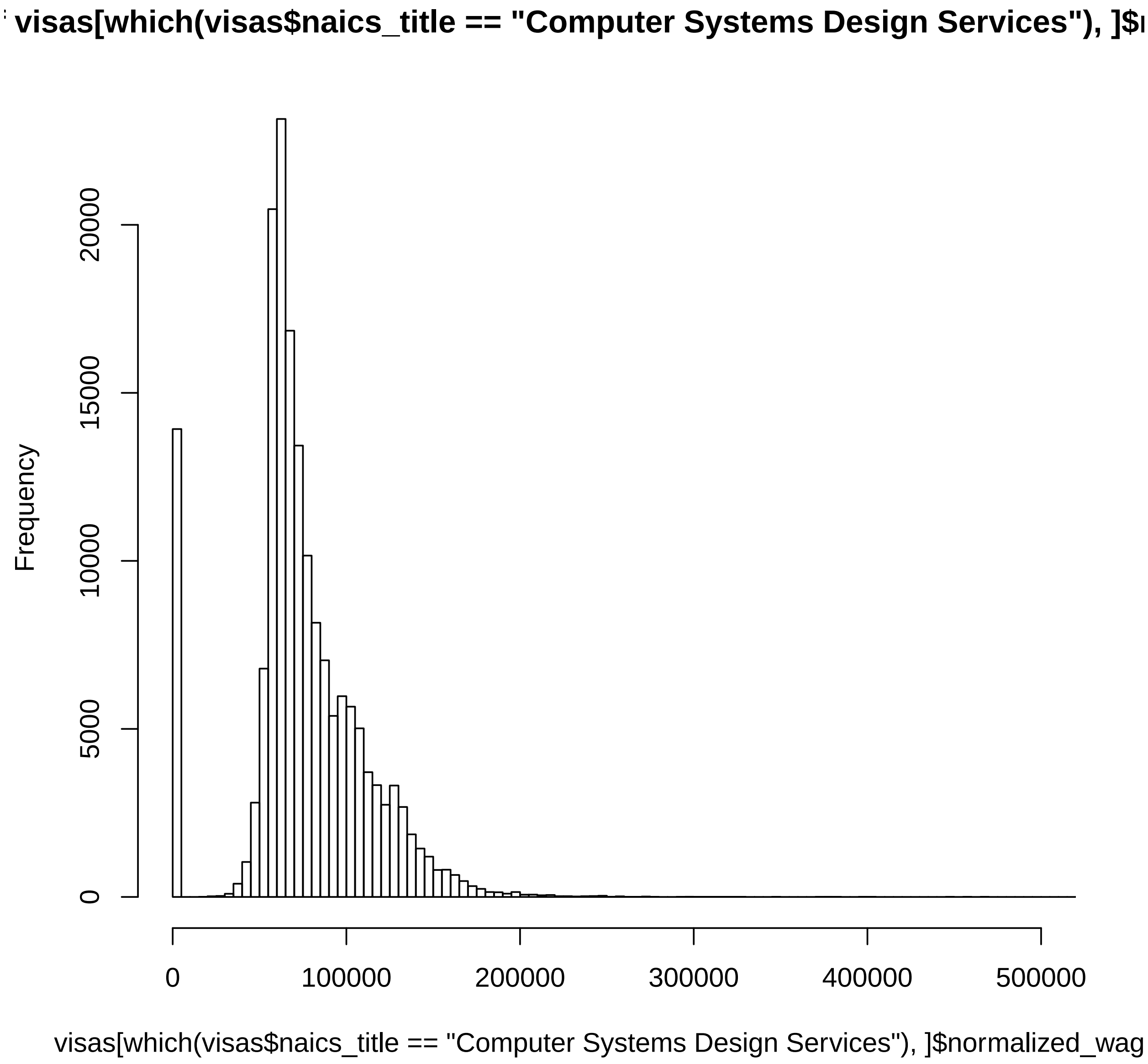
Once we remove all requests for more than 1,000 workers, we can start to see a pattern. The number of denied and withdrawn applications remains fairly constant, but the number of certified workers shows a steady rise after 2010, most likely due to a strengthening economy and returning jobs.

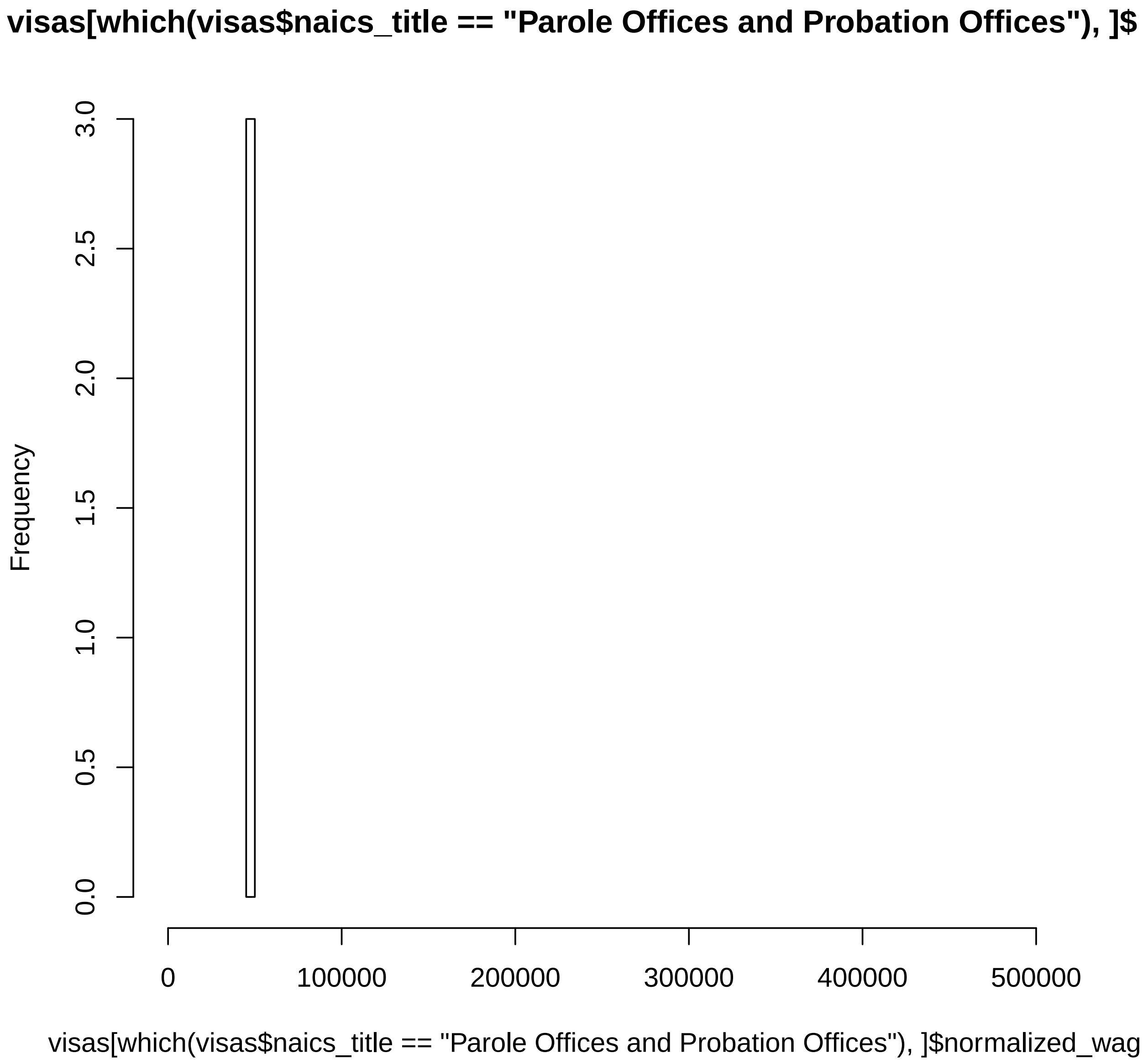
Table 1: Applications requesting over 1000 workers

fy	total_workers	status
2011	2010	DENIED
2012	2011	DENIED
2013	2012	DENIED
2013	2012	DENIED
2014	2013	DENIED
2011	10000	DENIED
2012	43555	DENIED
2010	52000	DENIED
2010	53000	DENIED
2010	53000	DENIED
2013	58500	DENIED
2013	60000	WITHDRAWN
2013	793172	DENIED
2011	1132014	DENIED
2013	2031308	DENIED
2010	7000000	DENIED




Number of workers by application status vs. U.S. job market, 2008 − 2016
2008
2009
2010
2011
2012
2013
2014
2015
2016
2000000
0
−2000000
−4000000
tot
status
CERTIFIED
JOBS
Certified H−1Bs vs. U.S. Job Market


visas[which(visas$naics_title == "Computer Systems Design Services"),]$
Frequency
0
5000
10000
15000
20000
0
100000
200000
300000
400000
500000
visas[which(visas$naics_title == "Computer Systems Design Services"),]$normalized_wag



4.3 Offered Wage

4.4 H-1B vs. PERM

5 Conclusion

```
#visas <- readRDS('H1BVisas.rds')
#perm <- readRDS('PermData.rds')
#perm.map <- readRDS('PermEmpMapsdat.rds')

#hist(visas[which(visas£normalized_wage < 250000),]£normalized_wage, breaks=500)

#findmode <- function(x, na.rm = TRUE) {

  #if(na.rm){
    #x = x[!is.na(x)]
  #}

  #ux <- unique(x)
  #return(ux[which.max(tabulate(match(x, ux)))])
#}

#wage.mode <- findmode(visas£normalized_wage)
#wage.mode
#abline(v=wage.mode + 500, col="red")

#hist(visas[which(visas£normalized_prevailing_wage < 250000),]£normalized_prevailing_wage, breaks=500)
#pw.mode <- findmode(visas£normalized_prevailing_wage)
#pw.mode
#abline(v=wage.mode + 500, col="red")
```

References

[1] U.S. Department of Labor, Office of Foreign Labor Certification Disclosure Data, https://www.foreignlaborcert.doleta.gov/performancedata.cfm