MA415/615 Project One: U.S. Total Illegal Alien Apprehensions Analysis

Xi Cheng, Ziran Min, Shih-Ching Huang

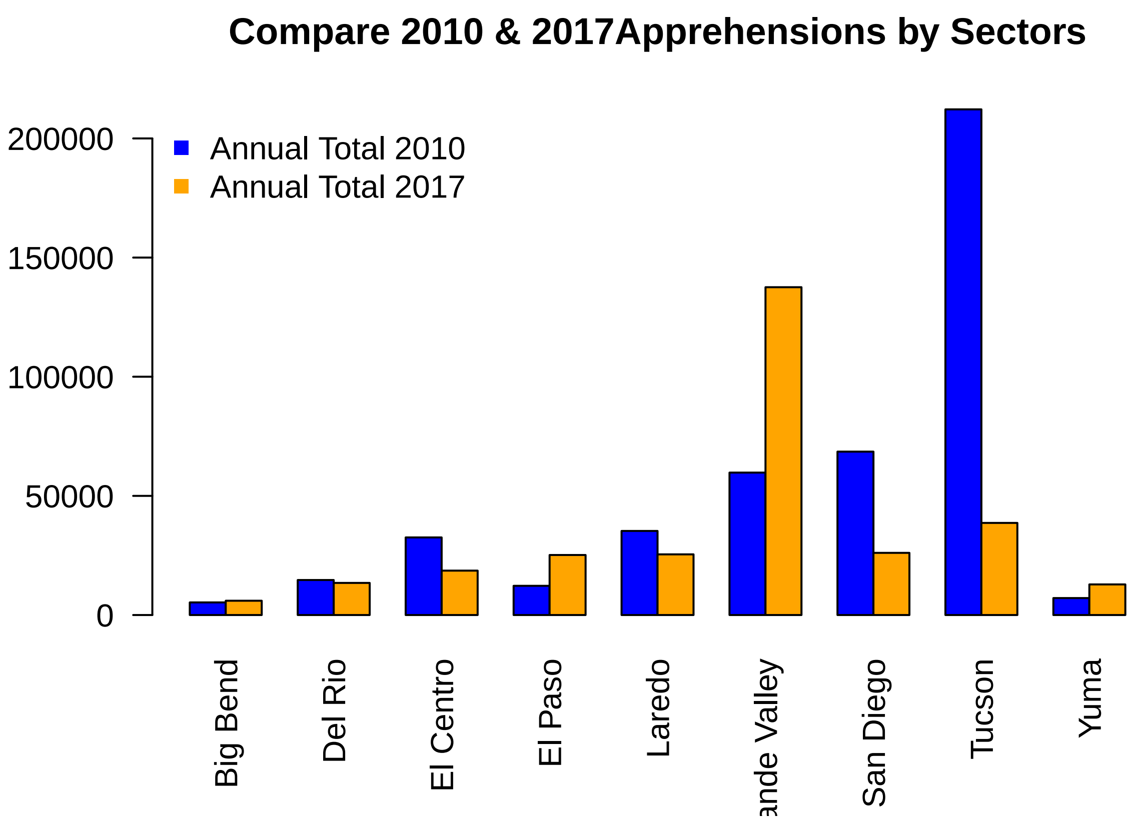
February 18, 2018

Background

Based on CNN news, we realized that the US-Mexico broader apprehensions hit 17 years lowest in the first few months of the Trump administration. This phenomenon was interpreted as "President Donald Trump's rhetoric and aggressive push to enforce immigration laws". Therefore, we want to analyze the data from the past to see whether this phenomenon had already existed before or it's really related to the new law as people saying.

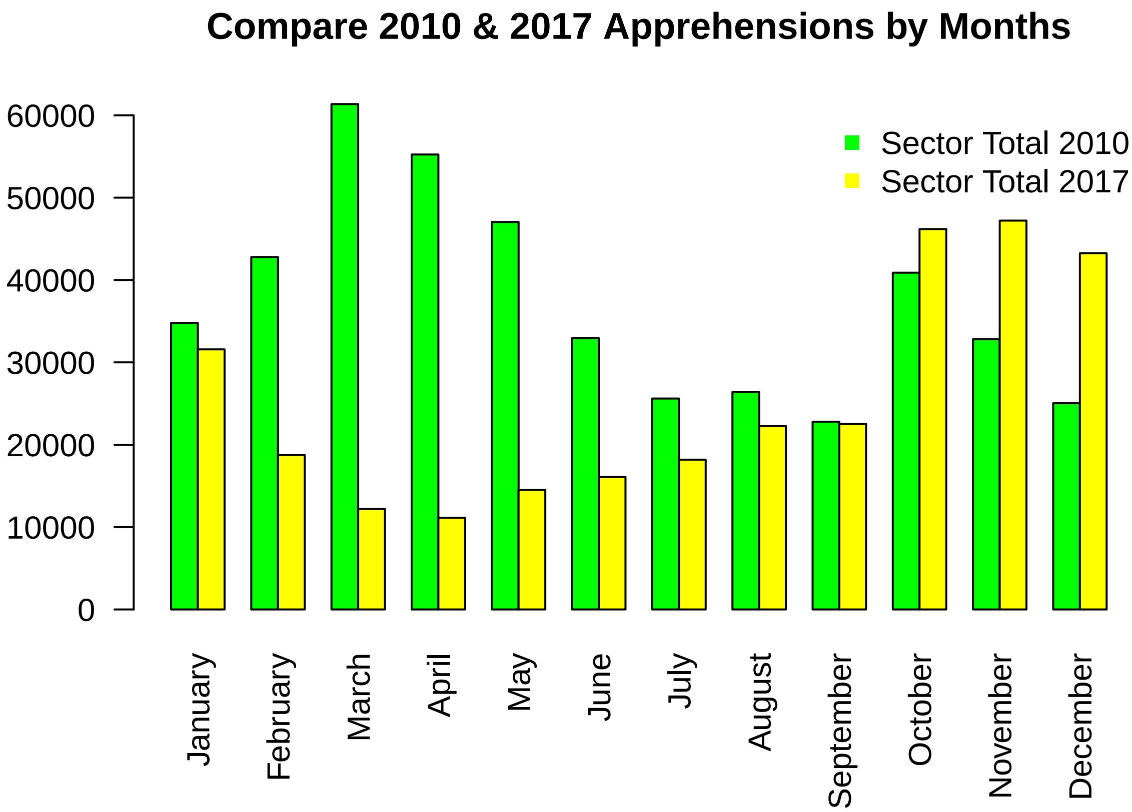
Data Visualization

First, we can visualize the data we have already got in the file as below. To compare the apprehensions by sectors, we first summed the apprehensions of each sector through 12 months and built the following plot:



We found that in 2010, Tucson has the most number of apprehensions, and in 2017 the sector with most apprehensions is Rio Grande Valley. But has there been a change in maximum from 2010 to 2017? We will focus on this question in next section.

To compare apprehension by months, we summed the apprehensions of each month through sectors and built the following plot:



Then, we found out that in March 2010 has the most apprehensions, followed by April and May, and in 2017, the most apprehensions changed to November, which then followed by October and December. We will also go deeper about whether the mean of the 3-month highest apprehensions changed later.

We also built shiny website https://chengxi.shinyapps.io/MA415-615-Project-One-Ziran-Xi-ShihChing/ to demonstrate more detailed plots to compare apprehensions across different sectors and months.

Using t-test to Solve Previous Questions

To find and compare the sector with the most apprehensions in 2010 and with the sector with the most apprehensions in 2017. We ran the following code:

```
#find the index of the sector that has most apprehensions in 2010
most_apprehensions_2010 <- match(max(B2010[1:9,13]), B2010[1:9,13])
#it is Tucson
rownames(B2010)[most_apprehensions_2010]
```

```
## [1] "Tucson"
```

```
#find the index of the sector that has most apprehensions in 2017
most_apprehensions_2017 <- match(max(B2017[1:9,13]), B2017[1:9,13])
#it is Rio Grande Valley
rownames(B2017)[most_apprehensions_2017]
```

```
## [1] "Rio Grande Valley"
```

After that, we ensured the most apprehensions in 2010 & 2017 is Tucson and Rio Grande Valley. To whether they have the mean, we conduct the following hypothesis test:

$$H_0 : \mu_{Tucson,2010} - \mu_{GRV,2017} = 0$$
$$H_a : \mu_{Tucson,2010} - \mu_{GRV,2017} \neq 0$$

```
##
##  Welch Two Sample t-test
##
## data:  B2010_Tuson and B2017_RGV
## t = 1.9547, df = 21.973, p-value = 0.06346
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -379.5935 12819.5935
## sample estimates:
## mean of x mean of y
##   17683.5   11463.5
```

Through this Welch Two Sample t-test, we get p-value = $0.06346 > 0.05$, so we fail to reject the null hypothesis that the mean of apprehension of Tucson in 2010 is equal to the mean of apprehension of Rio Grande Valley in 2017, and therefore conclude that there is no significant difference between the maximum of apprehensions by secoter in 2010 and 2017

To solve the second question we came up with in previous section, whether the mean of the 3-month highest apprehensions changed, we did the similar hypothesis test as the following:

```r
#find the three months that has the most apprehensions in 2010
most_3m_apprehensions_2010 <- order(B2010[10,1:12], decreasing = TRUE)[1:3]
#they are March, April, and May
colnames(B2010)[most_3m_apprehensions_2010]
```

```
## [1] "March" "April" "May"
```

```r
#find the three months that has the most apprehensions in 2017
most_3m_apprehensions_2017 <- order(B2017[10,1:12], decreasing = TRUE)[1:3]
#they are November, October, and December
colnames(B2017)[most_3m_apprehensions_2017]
```

```
## [1] "November" "October"  "December"
```

Null hypothesis and alternative hypothesis:

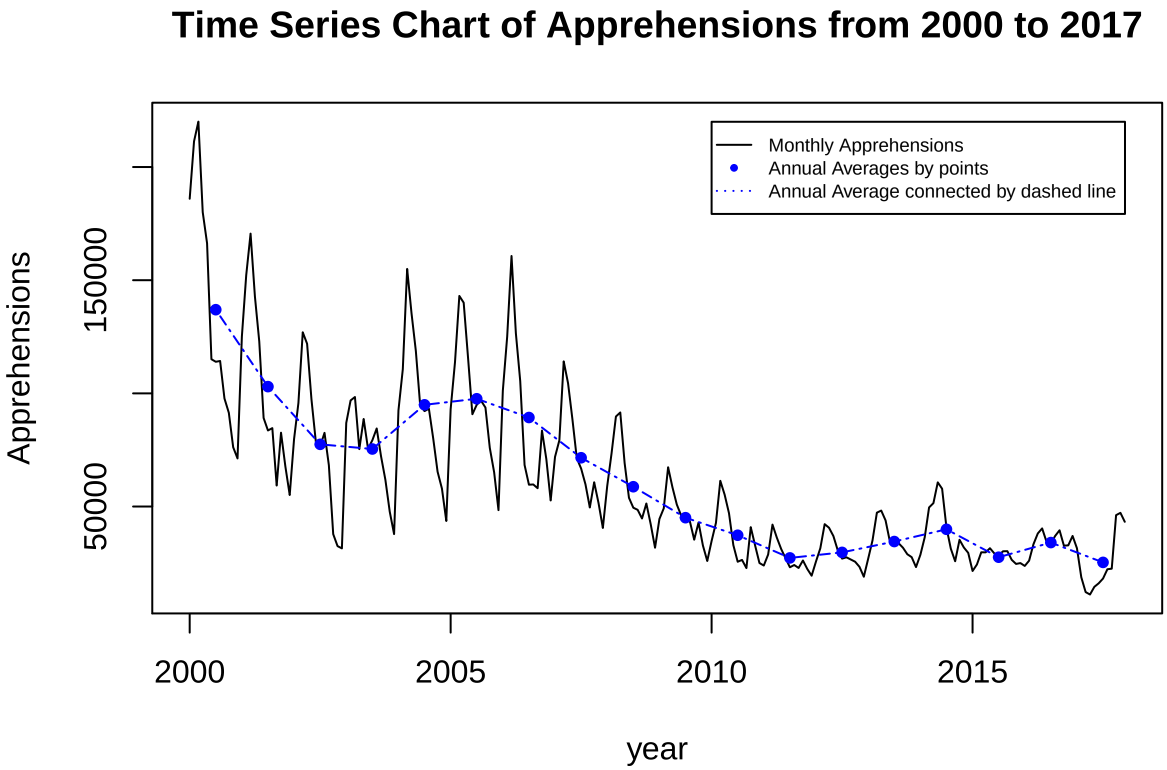
$$H_0 : \mu_{3months,2010} - \mu_{3months,2017} = 0$$

$$H_a : \mu_{3months,2010} - \mu_{3months,2017} \neq 0$$

```
##
##  Welch Two Sample t-test
##
## data:  B2010_3m and B2017_3m
## t = -5.9857, df = 11.05, p-value = 8.944e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -31591.20 -14611.46
## sample estimates:
## mean of x mean of y
##   2225.00  25326.33
```

Then we get p-value = 0.3454 > 0.05, so we fail to reject the null hypothesis that the mean of apprehension of from March to May in 2010 is equal to the mean of apprehension from October to December in 2017, and therefore conclude that there is no significant difference between the maximum of apprehensions from March to May in 2010 and from October to December in 2017.

Time Series



According to the Time series shown above, we can see that the apprehensions of America keep declining through 2000 to 2017, and reached 17 years lows in 2017.

Conclusion

The decreasing trend and the new lowest level we found from time series chart indeed reflect the effect of President Trump's and the US Department of Homeland Security's strict regulation on illegal apprehensions across the southern border. However, the t-tests we did don't show significant diference between the maximum levels of apprehensions through years. This is acceptable because we didn't compare the mimimum changes through years and our comparisons are across different sectors and months. To further study on the US illegal apprehensions in recently years, we need to conduct more detailed research how the apprehensions in each specific sector changes through months and years.