mapstu

Jian Lu

January 23, 2017

Abstract The **mapstu** package uses the geographical distribution data of the entire Williams College student population published in the Williams Course Catalogs to compare which states, and countries students are from each year. This is done for fifteen years from 2000 to 2015 by pulling data from text versions of the geographical distribution data. The core components of the package creates different choropleth maps based on the arguements passed to the function. With the visual support of the choropleth map, one can easily visualize the difference of student geopgraphical distribution between any two years of the dataset. A printed comparison of the two years can also produced.

Introduction

The geographical distribution of students from Williams College is an interesting dataset that can show the diversity of the student population. A visualization of the dataset can easily present the student diversity of the college, and the changes of where students are accepted from each year. The package **mapstu** pulls and cleans data from the yearly Williams College Course Catalogs and organizes the data into a readable csv file. Then the package reads the csv files into a dataframe and combines all the dataframes from all years merging into a complete dataframe which can be easily manipulated.

The vectors of the dataframe, or years, can be called on and manipulated into choropleth maps with the help of the S4 classes in R. There are two diffferent S4 datasets in the package. One includes all of the countries in the world, and the other includes all U.S. owned territories. The years can be combined with either of the S4 datasets, and then plotted. The package also includes another function which plots the change between any two years, and creates a plot where gradients of red and green represent decreases and increases.

Data

The data used to construct the overall dataframe in this package was taken from the Williams College Course Catalogs, from the website of the Office of The Registrar of Williams College, from years 2000 through 2015 by copy pasting the text directly into text files. These text files were then read into R, parsed through with the function **readYears**, and then manually edited to keep names identical across all years and to denote the difference between Georgia (State) and Georgia (Country). The clean versions of the files had the format (area, number of students) so that it could easily be read into R as a csv file. Thus it became possible to create a dataframe for each year from 2000 - 2015; however, it the data would be alot easier to manipulate if everything was merged into one dataframe. Thus by using recursion, and the **merge** function in R, all of the year dataframes were compiled into a single dataframe by the function **totaldata**. The function also replaces any NA values generated by the **merge** function with 0 to make calculations easier when plotting.

```
yearsdata <- mapstu::totaldata()
yearsdata[1:6, c("State.Countries", "X2000", "X2001")]
```

```
##   State.Countries X2000 X2001
## 1         Alabama     6     3
## 2          Alaska     5     8
## 3       Argentina     1     1
## 4         Arizona     5     5
## 5        Arkansas     1     3
## 6         Austria     2     2
```

Usmap

The function **usmap** allows for the visual representation of the change in geographic distribution for U.S. territories. This function uses a shape file of all U.S. territories and appends the difference of the year vectors we want to compare, and maps the data to a red/green color scheme. Gradients of red represent a decrease and gradients of green represent an increase in the student population at Williams.

An example here shows the change from William's student geographical distribution in 2000 to the William's student geographical distribution in 2015.

```
mapstu::usmap(yearsdata$X2015, yearsdata$X2000, title = "Change in Students 2000-2015")
```

```
## Warning: package 'tmap' was built under R version 3.3.2
```

```
## Warning: package 'tmaptools' was built under R version 3.3.2
```

We can immediately see some drastic changes in the geographical distribution of Williams College in these past 15 years. The student population from California has drastically increased, whereas the student population from Wisconsin has decreased. The downside is that the plot is somewhat small, and hard to see the individual state borders. Thus there is an *interactive* agrument in this function to open an interactive mode in the viewer.

```
mapstu::usmap(yearsdata$X2015, yearsdata$X2000, title = "Change in Students 2000-2015",
    save = TRUE, interactive = TRUE)
```

By running the code above, we also save the plot under the name of the title and open an interactive mode where one can zoom in and out. As well the interactive mode labels each state, and clicking on each area results in the numerical change of students in that area to be shown.

Countrymap

Worldplot

Sidebyside

Comparison

Perchange

Data analysis

Conclusion

Appendix