mapstu

Jian Lu '19

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 cloropleth maps based on the arguments passed to the function. With the visual support of the cloropleth map, one can easily visualize the difference of student geographical distribution between any two years of the data set. A printed comparison of the two years can also produced.

Introduction

The geographical distribution of students from Williams College is an interesting data set that can show the diversity of the student population. A visualization of the data set 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 data frame and combines all the data frames from all years merging into a complete data frame which can be easily manipulated.

The vectors of the data frame, or years, can be called on and manipulated into cloropleth maps with the help of the S4 classes in R. There are two different S4 data sets 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 data sets, 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 data frame 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 data frame for each year from 2000 - 2015; however, the data would be much easier to manipulate if everything was merged into one data frame. Thus by using recursion, and the **merge** function in R, all of the year data frames were compiled into a single data frame 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
```

Note In order to utilize the functions in this package, the dataframe must be initialized as the packages take in vectors from the yearsdata dataframe.

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")
```

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* argument in this function to open an interactive mode in the viewer. Example code of how to enter the arguments is:

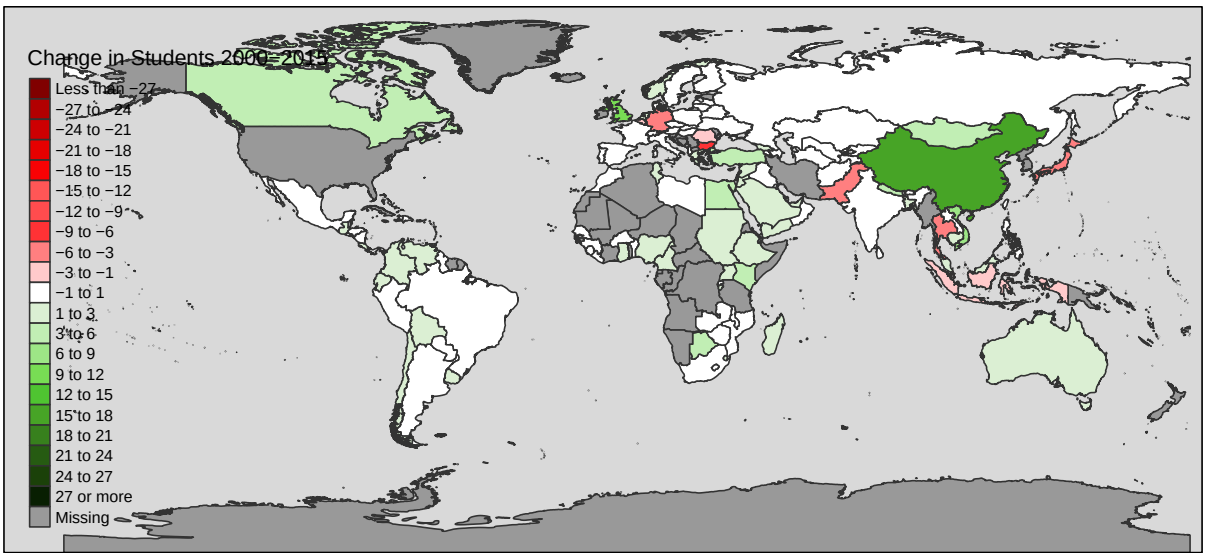
```
mapstu::usmap(yearsdata$X2015, yearsdata$X2000, title = "Change in Students 2000-2015",
    save = TRUE, interactive = TRUE)
```

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

The function **countrymap** is similar to **usmap**, but instead creates a visual representation of the change in geographical distribution for the international students at Williams. An example of the function showing the change between the years 2000 to 2015 is shown below.

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



A major downside to both **countrymap** and **usmap** is that the change has to be separated by domestic and international students, since the S4 files used do not have the states of the U.S. mapped with a world map. A future modification may also to display the overall change of the U.S. on the international plot.

Worldplot

The function **worldplot** is used to just map a single vector year of the geographical distribution of students. This is to show the overall diversity of the student population at Williams. An example of the function with the year 2015 is:

```
mapstu::worldplot(yearsdata$X2015, title = "Students Distribution in 2015")
```

Sidebyside

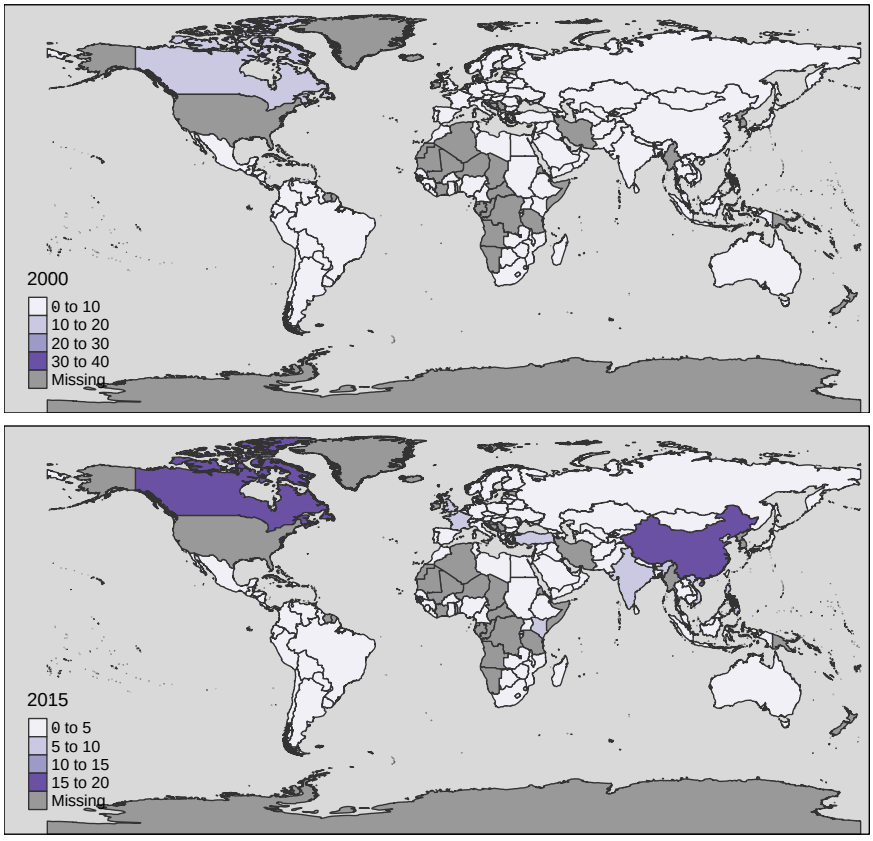
The function **sidebyside** is used to create two visuals of the above functions on the same plot. It can be used to directly compare different student population distributions from different years, or to compare changes in student distributions over two different time periods. Examples of different kinds of usage of the function are shown below:

Showing the geographical distribution in 2000 VS. 2015 in the U.S.:

```
mapstu::sidebyside(yearsdata$X2000, yearsdata$X2015, title1 = "2000",
    title2 = "2015")
```

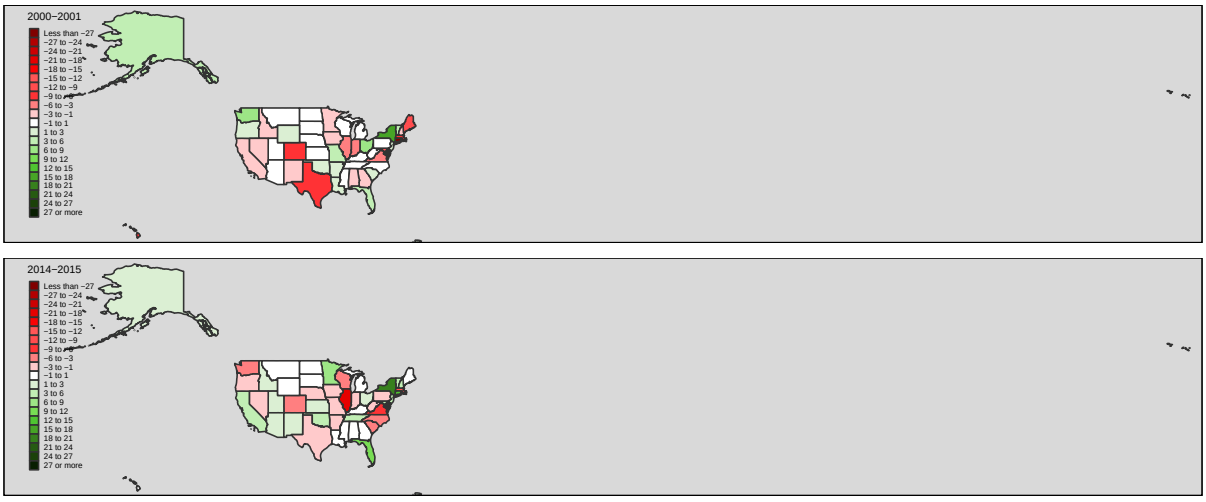
Showing the geographical distribution in 2000 VS. 2015 on the international level:

```
mapstu::sidebyside(yearsdata$X2000, yearsdata$X2015, title1 = "2000",
    title2 = "2015", US = FALSE)
```



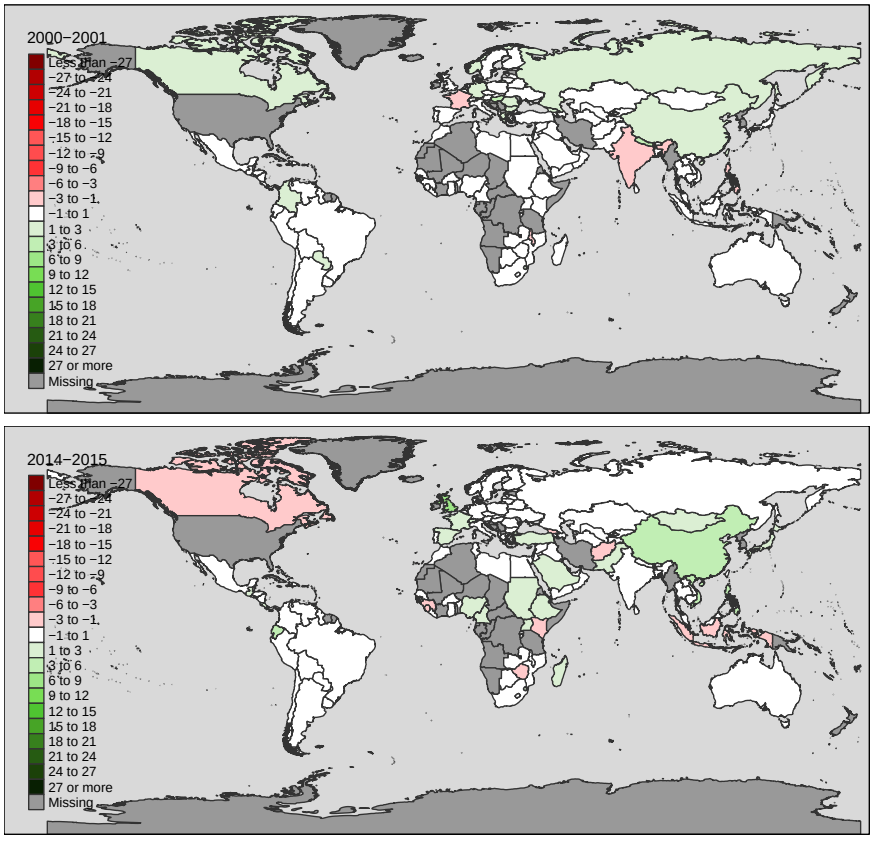
Showing the change in geographical distribution from 2000-2001 VS. 2014-2015 in the U.S. :

```
mapstu::sidebyside(yearsdata$X2001 - yearsdata$X2000, yearsdata$X2015 -
    yearsdata$X2014, title1 = "2000-2001", title2 = "2014-2015",
    change = TRUE)
```



Showing the change in geographical distribution from 2000-2001 VS. 2014-2015 on the international level :

```
mapstu::sidebyside(yearsdata$X2001 - yearsdata$X2000, yearsdata$X2015 -
    yearsdata$X2014, title1 = "2000-2001", title2 = "2014-2015",
    US = FALSE, change = TRUE)
```



Note: When using **sidebyside** to compare change, be sure to input **change = TRUE** argument.

Comparison

This function **comparison** takes in two different year vectors and returns a quick summary in a data frame comparing both years. Utilization of this function can provide actual numerical comparison between years rather than just a visual plot. Using **comparison** and **sidebyside** together will provide data for easier analysis. An example of the comparison between year 2000 and 2015 is shown below:

```
mapstu::comparison(yearsdata$X2015, yearsdata$X2000)
```

```
## [1] "Max Change:"
##          NAME Change
## 12 California    111
## [1] "Min Change:"
##             NAME Change
## 48 Massachusetts    -61
```

```
##                                 V1         Year1 Year2
## 1      Number of Domestic Students 2069.00000000  2094
## 2 Number of International Students  203.00000000   185
## 3         Total Number of Students 2272.00000000  2279
## 4                       Mean Change   -0.03381643    NA
```

Note: The more current year is put into the year1 argument, and the older year is put into the year2 argument.

Perchange

The function **perchange** takes in two different year vectors and returns a data frame with percent change calculated for each state and country. Because it is hard to represent Inf values and NA values visually, all of the Inf and NA values were changed to 0. The percent change data is very helpful to see how Williams has changed their acceptance in different areas over time. Since it is returned in a data frame, the percentage change can also be graphed in any of the graphing functions above; however, the plots will be inaccurate as Inf was set equal to NA values. An example of **perchange** used on year 2000 and 2015.

```
perch <- mapstu::perchange(yearsdata$X2015, yearsdata$X2000)
perch[1:6, c("NAME", "Perchange")]
```

```
##        NAME Perchange
## 1   Alabama       -50
## 2    Alaska         0
## 3 Argentina         0
## 4   Arizona        80
## 5  Arkansas       200
## 6   Austria       -50
```

Conclusion

A comparison between the geographical distribution of students attending Williams in the year 2000 to the students attending Williams in the year 2015 show that there has been a 18 person increase in the number of international students. During this time period Williams' largest increase in distribution lies in the 111

student increase from California state. The largest decrease in distributions lies in the 61 student decrease from Massachusetts state. The graphics produced above indicate that Williams College has definitely moved towards increasing the amount of student diversity of different geographical areas. In the sidebyside plot of the years 2000vs2015, there is an obvious and large increase in the density of international students at Williams College, especially for students in Asia and Canada. It can also be seen that there are more students coming from Europe as well. As well, the "Change in Students 2000-2015" plot shows some major shifts in the geographical distribution of students within the U.S. These visual plots and numerical comparisons point towards Williams College attempting to diversify its student geographic distribution away from the northeastern region of the U.S. It seems that Williams College is seeking for greater student diversity within the college, and also attempting to spread its name away from the typical northeastern region of NESCAC schools in the U.S. However, these are merely hypotheses and cannot be proven without further investigation.

The package **mapstu** includes functions that allow for easy plotting of the geographical distribution data found in the Williams College Course Catalogs. It allows for easy manipulation of the years vector data so that the user can easily create plots of their choice which can be very helpful in studying the geographical diversity on campus, and how that changes throughout the years. However, the package can certainly have many improvements. The function **readYears** can be improved so that no manual editing of the names of countries is required. All the of the functions could also be formatted into Hadley Wickham's recommended **function**(dataset, x1, x2, ...) format. The plots could also be improved by merging the S4 classes of international countries and the states, resulting in a single plot, so multiple functions to plot the same data on two different plots would no longer be required. As well, it would be interesting to find data on student cultures, or ethnicities and map them on them on top of the plots. This would allow for us to draw conclusions not just about geographical diversity, but about how geographical diversity can affect the cultural diversity of a campus.

Appendix

The geographical distribution data throughout the years found in the Williams Course Catalogs were often in the form of (Alabama ..... 5). The text from these pdfs were copy-pasted over into a text file and renamed by their years. In order to create readable csv files, all extra punctuation was removed from these text tiles (Alabama 5). Since the removal of extra punctuation left spaces all extra spaces were removed (Alabama5). Then a comma was inserted between any lowercase letter and a number (Alabama,5). In the cases where areas had two word names, spaces were inputted between any lowercase and uppercase letter (New York,3). Finally to separate the states with their values spaces were inputted between any numerical value and a letter (Alabama,5 New York,3).