MA School Districts Performance Analysis

Cesar O Gonzalez Tellez. S0315654

**Prof. Marcos Luna, Ph. D.**

I am interested in Education around the state of Massachusetts. I am concerned that several towns have poor performance compared with their neighbors, sometimes considerably.

I decided to check if the investment per pupil is a significant factor in the performance of the education departments of our towns. Originally, I wanted to use MCAS results compared against **expenditure per pupil** in Massachusetts. However, the MCAS data was very inconsistent year after year, so I decided to go for a better option. I found that the **graduation rate at High School level** is more consistent and more reflective of the reality in every school district, so I decided to compare it with the expenditure.

I use global variables that allow me to change the year of the study and the working directory. This way, I can run the analysis on several years in order to identify issues, environments, and other factors.

I also made a function to trim strings. I use it when dealing with the data in CSV mode.

# Global Variables and functions to use. Adapt these variables for your environment  
# We will filter by year.   
theYear <- 2016  
# The working directory with the maps and files  
theWD <- "c:/gisclass/gph942/final"  
# This function returns string w/o leading or trailing whitespace  
trim <- function (x) gsub("^\\s+|\\s+$", "", x)

The script reads a map previously dowloaded from The Massachusetts GIS website (<https://www.mass.gov/service-details/massgis-data-layers>).

I imported the tables with data from Massachusetts Department of Elementary and Secondary Education (<http://profiles.doe.mass.edu/>). They are Excel files with no geographic information other than the name and ID of the school district. The tables are converted as CSV format for easier management.

Reading the data tables

gradrates <- read.csv(file.choose(),stringsAsFactors = FALSE,header=TRUE)  
expenditures <- read.csv(file.choose(),stringsAsFactors = FALSE,header=TRUE)

**The data from these two sources groups several towns in one single school district**. This is so unfortunate because the maps generated have many towns missing. There could be several ways to try to fix this situation, but they are out the original scope of this project.

#### DATA WRANGLING

By glimpsing the tables I found that the fields OrgCode and DistrictID --from the gradrates and expenditures tables respectively-- represent the id equivalent to town id in the spatial map. But the numbers are in the 10,000s. I need to divide this field by 10,000 in order to get the correct id value as string.

Also, I need to remove the '$' and ',' characters from the Expenditure per pupil field. Then convert it to numeric.

gradrates <- gradrates %>%  
 filter(Year==theYear) %>%  
 mutate(townidG=as.character(OrgCode/10000))

## Warning: package 'bindrcpp' was built under R version 3.4.2

expenditures <- expenditures %>%  
 filter(Year==theYear) %>%  
 mutate(townidX=as.character(DistrictID/10000))  
  
expenditures$TotalExpendituresPerPupil <- trim(gsub('$','',gsub(',','',expenditures$TotalExpendituresPerPupil,fixed=TRUE),fixed=TRUE))  
expenditures$expenditure <- as.numeric(expenditures$TotalExpendituresPerPupil,na.rm=TRUE)

Let's figure out what's happening with this information

favstats(gradrates$PrctgGraduated)

## min Q1 median Q3 max mean sd n missing  
## 9.1 86.4 92.5 95.7 100 88.49863 13.48023 293 0

favstats(expenditures$expenditure)

## min Q1 median Q3 max mean sd n  
## 10970.84 13939.12 15018.19 17639.94 32291.23 16124.88 3493.202 322  
## missing  
## 0

#### READING THE MAP

By loading the MASS GIS Map, we found that we need the TOWN\_ID field to join tables with graduation rate and expenditures.

MACensus2010Towns <- readOGR(dsn=theWD,layer="CENSUS2010TOWNS\_POLY",stringsAsFactors = FALSE)

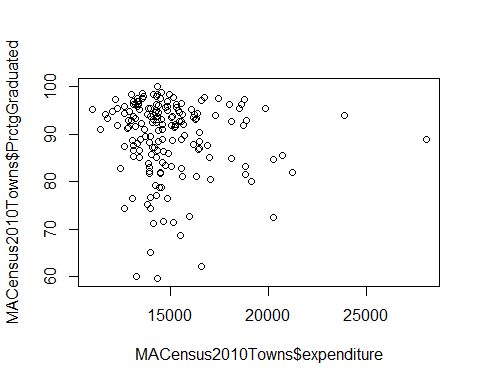
## OGR data source with driver: ESRI Shapefile   
## Source: "c:/gisclass/gph942/final", layer: "CENSUS2010TOWNS\_POLY"  
## with 351 features  
## It has 20 fields  
## Integer64 fields read as strings: POP1980 POP1990 POP2000 POP2010 POPCH80\_90 POPCH90\_00 POPCH00\_10 HU2010

MACensus2010Towns$TOWN\_ID <- as.character(MACensus2010Towns$TOWN\_ID)  
  
# Let's join the tables with the town id’s and the grad rate and spending fields   
MACensus2010Towns <- merge(MACensus2010Towns,gradrates[,c("townidG","PrctgGraduated")],  
 by.x="TOWN\_ID", by.y="townidG")  
MACensus2010Towns <- merge(MACensus2010Towns,expenditures[,c("townidX","expenditure")],  
 by.x="TOWN\_ID", by.y="townidX")  
# Clean up time. Remove data without grad rate or expenditure  
MACensus2010Towns <- subset(MACensus2010Towns,!is.na(MACensus2010Towns$PrctgGraduated))  
MACensus2010Towns <- subset(MACensus2010Towns,!is.na(MACensus2010Towns$expenditure))

#### BASIC ANALYSIS

Let's plot both new variables (gradrate and expenditure per pupil) to check if there is a relationship between them

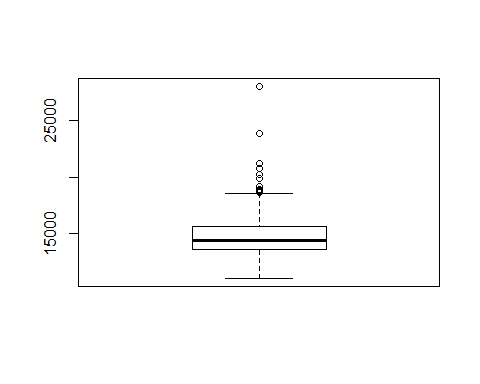
plot(MACensus2010Towns$expenditure,MACensus2010Towns$PrctgGraduated)



Most of the results are concentrated in one single range of spending and graduation rate.

What are the results with boxplotting? Let's start with expenditure per pupil

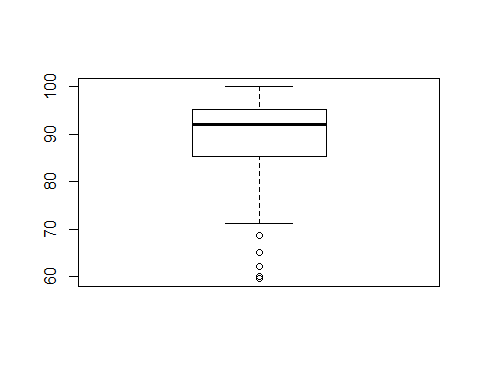
boxplot(MACensus2010Towns$expenditure)



The spending is around the $15,000, with several outliers

Now, what about the graduation rate

boxplot(MACensus2010Towns$PrctgGraduated)



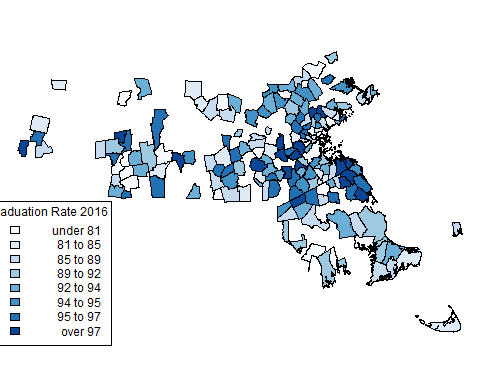
Shows what the favstats told us: median/mean in 92.5%/88.49%

We can see there is a not a big relationship between investment per pupil and graduation rates.

#### GEAGRAPHICAL RELATIONS

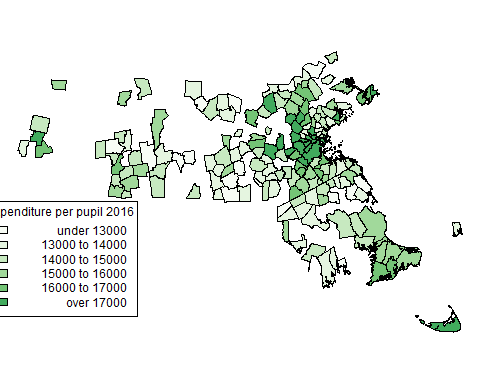
Let's start with a choropleth of graduation rate

par(mar=c(0,0,0,0)) # Set the minimum margins for the plots  
shades <- auto.shading(MACensus2010Towns$PrctgGraduated,n=8,cols=brewer.pal(8,"Blues"))  
choropleth(MACensus2010Towns,MACensus2010Towns$PrctgGraduated,shades)  
choro.legend(19445,862472,shades,title=paste("Graduation Rate",theYear),cex=0.75)



And now, let's see what's happening with expenditures

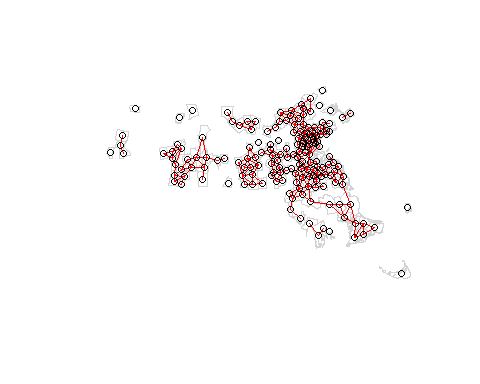
par(mar=c(0,0,0,0)) # Set the minimum margins for the plots  
shades2 <- auto.shading(MACensus2010Towns$expenditure,n=8,cols=brewer.pal(8,"Greens"))  
choropleth(MACensus2010Towns,MACensus2010Towns$expenditure,shades2)  
choro.legend(19445,862472,shades2,title=paste("Expenditure per pupil", theYear), cex=0.75)



In both cases, we see certain clusterings in some parts of the state. But that's for the independent variables by themselves.

Let's test for spatial autocorrelation. And check if there are similarities to each town's neghbors

matowns.nb <- poly2nb(MACensus2010Towns) # Get the neighbors  
plot(MACensus2010Towns,border="lightgrey")  
plot(matowns.nb,col="red",coordinates(MACensus2010Towns),add=TRUE)



#### GRADUATION RATE ANALYSIS

Let's look for clusters with the Moran plot

dev.off() # Restart device

## null device   
## 1

moran.plot(sub\_ma$PrctgGraduated,sub\_ma.lw)

Each town weight, compared against the average 5 neighbors, the line timidly suggests that every town is like their neighbors, but it is not of high relevance.

Let's perfom a Moran test to see if there is randomness in the results or if there is a pattern

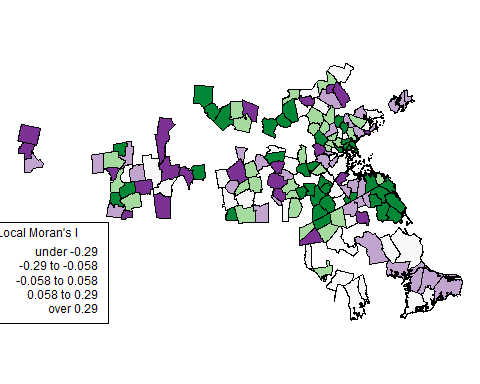
moran.test(sub\_ma$PrctgGraduated,sub\_ma.lw)

##   
## Moran I test under randomisation  
##   
## data: sub\_ma$PrctgGraduated   
## weights: sub\_ma.lw   
##   
## Moran I statistic standard deviate = 2.0102, p-value = 0.0222  
## alternative hypothesis: greater  
## sample estimates:  
## Moran I statistic Expectation Variance   
## 0.118851443 -0.006329114 0.003877715

Does data reflect correlation? p-value < 0.05... and it is , so our data is different than random, there are some clusterings; therefore, data is not random and there is a pattern.

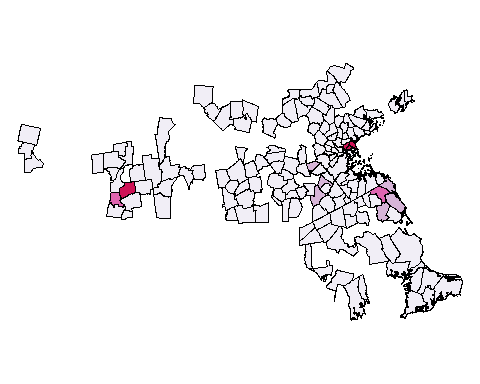
Where in MA are the clusterings? Let's use LISA

sub\_ma.li <- localmoran(sub\_ma$PrctgGraduated,sub\_ma.lw)  
ur.shade <- auto.shading(c(sub\_ma.li[,1],-sub\_ma.li[,1]),cols=brewer.pal(5,"PRGn"))  
par(mar=c(0,0,0,0))  
choropleth(sub\_ma,sub\_ma.li[,1],ur.shade)  
choro.legend(19445,862472,ur.shade,title="Local Moran's I",cex=0.75)



Purple divergents, green = clustering... Is not uniform... is not the same across the state! Therefore, we can assume that this is a complex model, with relations but hard to explain.

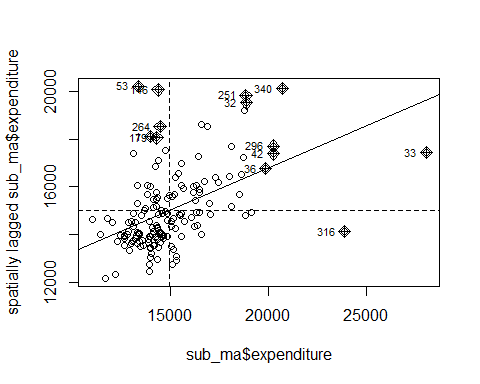
par(mar=c(0,0,0,0))  
pval.shade <- shading(c(0.01,0.05,0.1),cols=rev(brewer.pal(4,"PuRd")))   
choropleth(sub\_ma,sub\_ma.li[,5],shading = pval.shade) # the least significant statitics there is



There are just a few parts the state where the local moran is statisticaly significant

What about basing our analysis on expenditure per pupil?

moran.plot(sub\_ma$expenditure,sub\_ma.lw)



In this case, the line suggest a strong spatial auto-correlation, meaning that the spending per student is very similar among neighbors

Let's perfom a Moran test to see if there is randomness in the results or if there is a pattern

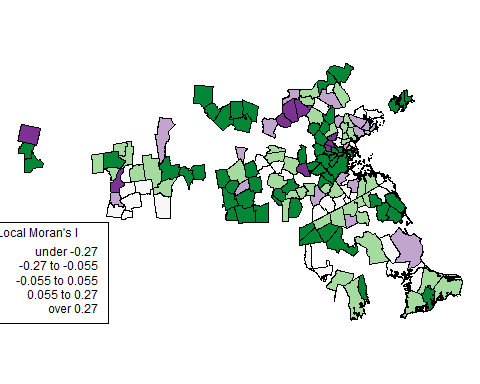
moran.test(sub\_ma$expenditure,sub\_ma.lw)

##   
## Moran I test under randomisation  
##   
## data: sub\_ma$expenditure   
## weights: sub\_ma.lw   
##   
## Moran I statistic standard deviate = 5.8709, p-value = 2.167e-09  
## alternative hypothesis: greater  
## sample estimates:  
## Moran I statistic Expectation Variance   
## 0.352112658 -0.006329114 0.003727538

Does data reflect correlation? Again, the p-value is extremely low, so our data is clearly not random at all. As with the graduation rate, there is some clustering; data is not random and there is a pattern.

Let's use LISA to see where in the state is the clustering

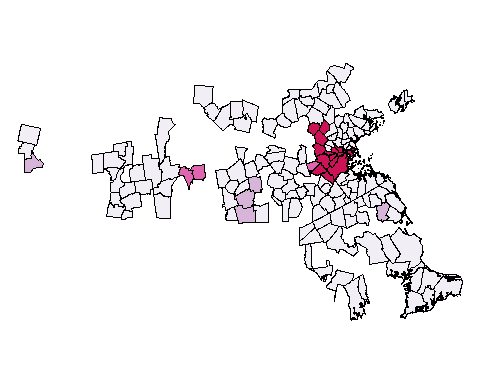
sub\_ma.li <- localmoran(sub\_ma$expenditure,sub\_ma.lw)  
ur.shade <- auto.shading(c(sub\_ma.li[,1],-sub\_ma.li[,1]),cols=brewer.pal(5,"PRGn"))  
par(mar=c(0,0,0,0))  
choropleth(sub\_ma,sub\_ma.li[,1],ur.shade)  
choro.legend(19445,862472,ur.shade,title="Local Moran's I",cex=0.75)



Purple divergents, green = clustering... Is not uniform... is not the same across the state! Once again, this reflects a complex model

Finally, create a pval shading scheme and plot it.

par(mar=c(0,0,0,0))  
pval.shade <- shading(c(0.01,0.05,0.1),cols=rev(brewer.pal(4,"PuRd")))   
choropleth(sub\_ma,sub\_ma.li[,5],shading = pval.shade)



The red areas that meet or exceed the key probability values of correlation among neighbors.

#### CONCLUSIONS

What does all this mean? Although we can see strong clustering of the independet variables --graduation rate and expenditure per pupil-- I could not find a clear correlation between the two variables or if one is a strong factor for the other one.

The region that I am most interested in is the Merrimack Valley of Massachusetts, especially in the towns of Methuen, Lawrence, Andover, and North Andover. Of them all, Lawrence represents an interesting case because in 2011, the state took over the administration of the town and took several initiatives in order to improve the performance of the schools. Among those changes were: replacement of personnel in all levels, implementation of new education programs, and increase in expending per pupil in order to accommodate new after school activities, STEM programs, and improvement of facilities.



The results are surprising. Out of the 4 towns, the one with less expenditure per pupil is actually the only one that has a steady performance. North Andover maintains a 94% graduation rate every year, no matter being the town spending the less in every student.

When the Lawrence education department was taken over by the state, the graduation rate was 52%. 4 years later, the same variable jumped 20 percentile points. This certainly means that the measures taken by the super-intendent's office worked effectively, but that involved an increase of money per every student.

Is more spending the key for success in public schools? There is not enough evidence to support this assertion. The case of North Andover is clear example of how a low budget has been used in favor of the students.

But, the schools of Lawrence clearly needed an influx of cash in order to improve their rates.

Maybe, one conclusion could state that, for failing districts once the necessary programs and policies are firmly established, and the measurable parameters are achieved, the flow of money will tend to stabilize.

Also, higher graduation rates do not necessarily mean better students, so more assessments are needed to claim a definite and final conclusion.

Further analysis, and the consideration of other variables, is necessary.