APSTA-GE 2011 Project #2: Clustering on Provinces in Mainland China according to Educational Conditions

Yuheng Ling (yl4042)[†] and Yuan Ding (yd1400)[†]

[†]Courant Institute of Mathematical Sciences, New York University
[††]College of Global Public Health, New York University

January 2019

1 Overview

1.1 Background

After 40 years of social reform and the opening up policy, China has achieved remarkable progress, not only in economic growth but also in education and etc. Despite the positive achievement, the inequality of education among different regions of China has always been one of the most severe obstacles that require the government to overcome. Even though the gap still remains, in recent years, the government gradually shifts its focus on equality educations issues. Meanwhile, in 2013, the Ministry of Education of the PRC launched the most important education reform opinions in recent 30 years, announcing that the emphasizing of education should be transferred to "Quality-Oriented Education". The opinions indicate that the education should break the constriction of merely learning in class curriculum, and cultivating student's interesting in extracurricular activities.

1.2 Objective

Our objective is to combine these two aspects, which are "the education inequality" and "Quality-Oriented Education", utilizing different methods for clustering analysis for the province.

- Separating the provinces to different categories based on the progress of transferring to "Quality-Oriented Education".
- Exploring which provinces may have the difficulty and may need more support to achieve the goal.

2 Data Preparation

2.1 Data Source

Our data is extracted from two government websites' database:

- **the Ministry of Education of the PRC**: http://en.moe.gov.cn/documents/statistics/2013/national/[4]
- **National Bureau of Statistics of China**: http://www.stats.gov.cn/english/[2]

We group the data based on year and therefore form the dataset: **edu2013.csv**.

2.2 Features Exploration

Both of our two datasets have six variables for clustering: **HPTA13, GTHR13, EER13, HLR13, HSAR13, HMCR13** and **GC13**. The meanings of these six variables are listed in Table 1. In this project, we only care about variables which indicate the educational conditions of senior high schools, because primary and junior high school education are compulsory in China and there will be a large amount of financial support. We believe that an analysis focusing mainly on senior high schools can better reflect the educational conditions of a specific region.

Table 1: The Meaning of Raw Dataset Variables for the year 2013

Variable Name	Meaning
HPTA13	High School Pupil-Teacher Ratio
GTHR13	The ratio of high school teachers with master's degree
EER13	The ratio of education expenditure to the total expenditure
HLR13	The ratio of high schools with libraries
HSAR13	The ratio of high schools with sports areas
HMCR13	The ratio of high schools with multimedia classrooms

2.3 Additional Variables

In this project we will use three additional variables to make comparisons after clustering: Region (such as east, middle, west), Per capita GDP and Gini coefficient for education.

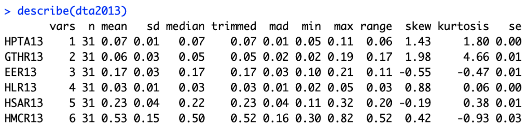
First, all the provinces in China are divided into three regional groups according to their geographical locations: **east, middle** and **west**. We want to discover if there is a correlation between location and educational condition. Second, we assume that for a specific province, the higher the Per capita GDP, the better the educational conditions. So the relationship between Per capita GDP and educational conditions is studied. Last, our dataset has an original variable **Gini coefficient for education**.[5] The higher this index is, the greater inequality of education is in that area. Therefore, we can also use this variable as a "measurement" for our clustering results.

3 Analysis Outline

3.1 Data pre-processing

3.1.1 Logarithm Transformation

Before we process our clustering analysis, we first summary our data and observe the distribution of each variable to check whether we need to do some pre-transformation. As we can see from Fig 1 & Fig 2, the distribution of **HPTA13** and **GTHR13** are badly right-skewed, which indicate that we should take logarithm transformation towards these two variables. After taking the log transformation, Fig 3 shows that the logarithm function did work for these two variables.



```
> describe(dta2013)
       vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
HPTA13    1 31 0.07 0.01   0.07    0.07 0.01 0.05 0.11  0.06  1.43     1.80 0.00
GTHR13    2 31 0.06 0.03   0.05    0.05 0.02 0.02 0.19  0.17  1.98     4.66 0.01
EER13     3 31 0.17 0.03   0.17    0.17 0.03 0.10 0.21  0.11 -0.55    -0.47 0.01
HLR13     4 31 0.03 0.01   0.03    0.03 0.01 0.02 0.05  0.03  0.88     0.06 0.00
HSAR13    5 31 0.23 0.04   0.22    0.23 0.04 0.11 0.32  0.20 -0.19     0.38 0.01
HMCR13    6 31 0.53 0.15   0.50    0.52 0.16 0.30 0.82  0.52  0.42    -0.93 0.03
```

Figure 1: Summary of Raw Data in 2013

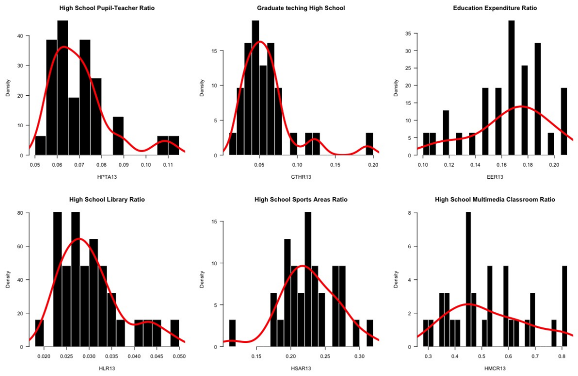


Figure 2: Histogram of Raw Data

3.1.2 Standardize Transformation

The logarithm transformation indeed optimizes the skewness, however, we find out that the scale of **HSAR13** and **HMCR13** remain a clear difference from other variables. Moreover, since our data mainly consists the ratio, most values approach 0, so we decide to standardize all variables. We did the PCA after standardizing the data as well, and it turns out that there exists some separation but not clear enough. Hence, the clustering analysis is necessary.

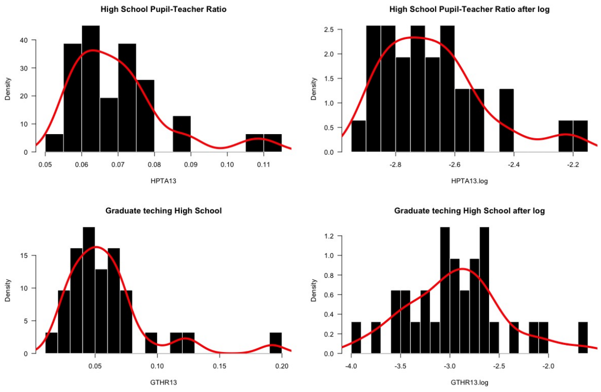


Figure 3: HPTA13 and GTHR13 after taking the log

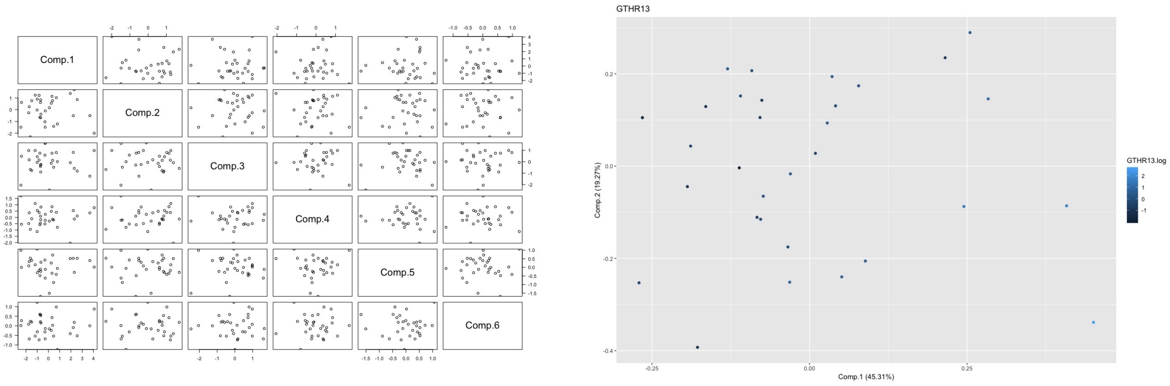


Figure 4: PCA after standardize

3.2 Clustering analysis

We will mainly focus on our clustering analysis based on two methods - Hierarchical Clustering and K-means Clustering, where several methods would be applied to determine the optimal clustering number. Apart from these two methods, we would also explore some other methods such as PAM Clustering to valid our outcome.

- Hierarchical Clustering
  - Cutree method
  - $C(g)$ method
- K-means Clustering
  - $C(g)$ method
  - Average silhouette method
  - Clustergram method
- PAM&Mixed Model Clustering

3.2.1 Hierarchical Clustering

First we use the hierarchical clustering, after trying all the three methods (see Figure 5) we find out that the complete method gives better performance on our dataset. The dendrograms of ”single” and ”centroid” methods indicate that there are too many small clusters, some of which can be very confused especially in the centroid dendrogram. The dendrogram of complete method looks much better: although each cluster is not of the same size, at least they are clearly separated from each other in a specific horizontal level. Therefore, we choose ”complete” method in our hierarchical clustering analysis.

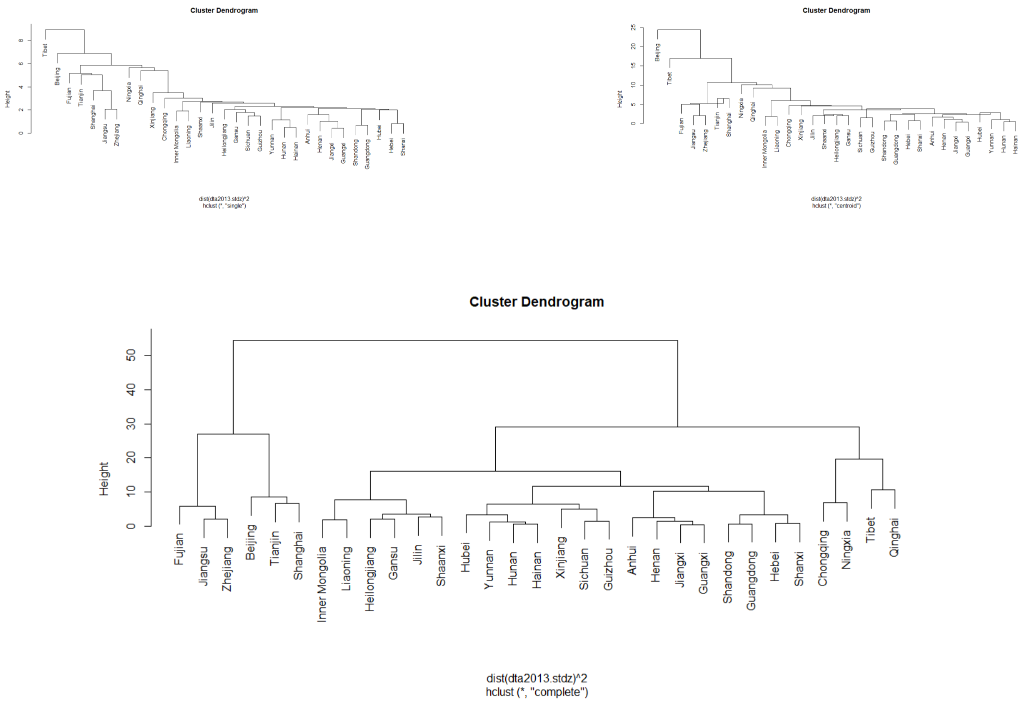
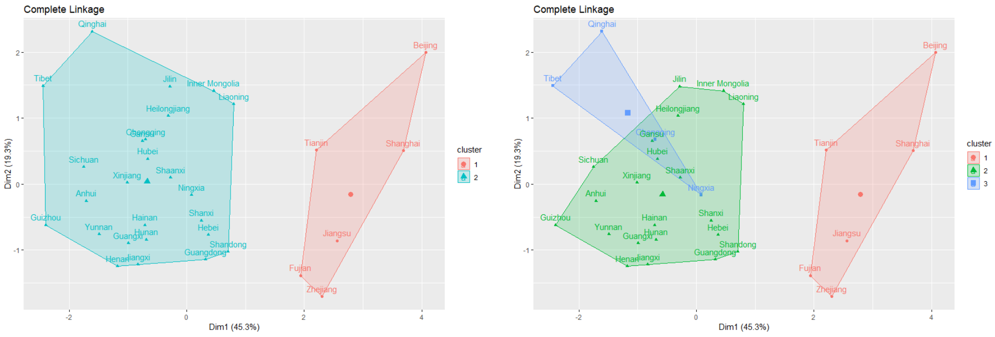


Figure 5: Dendrogram of Different Methods (hierarchical clustering)

Choose the number of clusters by Cutree method By examining the dendrogram of complete method, we can either choose the optimal number of clusters to be 2 or 3. When the number of clusters is 2 or 3, the points are pretty well separated. We draw both the plots using the cutree method to determine the number of clusters.

From Figure 6 we can see that the number of 2 would be a better choice because there is no overlap and two clusters are clearly separated from each other. When the number of clusters is 3, there is some overlap between two clustering groups.

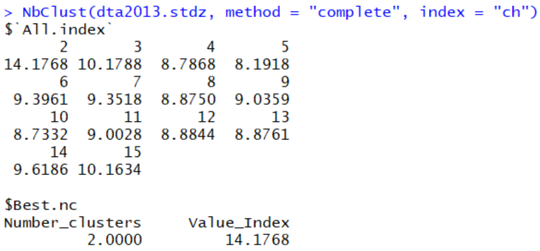
Determine the optimal number of clusters based on $C(g)$ Next we use the NbClust() function to determine the optimal number of clusters, which is mainly based on the C(g). The result of the NbClust also tells us that the optimal number of clusters in this case is 2 (Figure 7).



(a) n=2

(b) n=3

Figure 6: Assigned Cluster Labels (complete linkage)



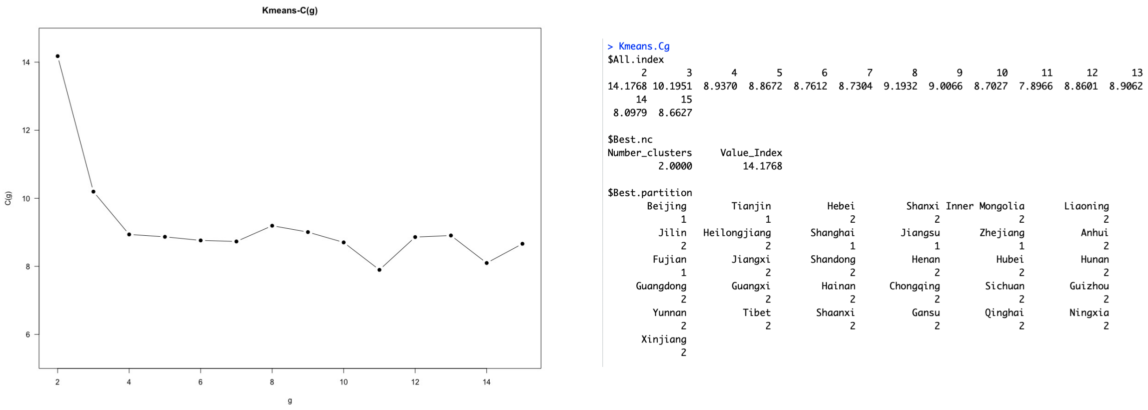
```
> NbClust(dta2013.stdz, method = "complete", index = "ch")
$`All.index`
      2       3       4       5
14.1768 10.1788  8.7868  8.1918
      6       7       8       9
 9.3961  9.3518  8.8750  9.0359
     10      11      12      13
 8.7332  9.0028  8.8844  8.8761
     14      15
 9.6186 10.1634

$Best.nc
Number_clusters     Value_Index
         2.0000         14.1768
```

Figure 7: The Result of C(g)

3.2.2 K-means Clustering

K-means Cluster has pretty high accuracy if we can determine the number of clusters in advance, so we use several methods and try to find out the number of clusters.

$C(g)$ method We also explore the $C(g)$ method in Kmeans clustering, and the figures show that $C(g)$ approaches the maximum when $g = 2$. Nevertheless, as we can see, $C(g)$ doesn't have so much difference when $g = 2$ or $g = 3$, so we would take both into our consideration.



```
> Kmeans.Cg
$All.index
      2       3       4       5       6       7       8       9      10      11      12      13
14.1768 10.1951  8.9370  8.8672  8.7612  8.7304  9.1932  9.0066  8.7027  7.8966  8.8601  8.9062
     14      15
 8.0979  8.6627

$Best.nc
Number_clusters     Value_Index
         2.0000         14.1768

$Best.partition
       Beijing        Tianjin          Hebei         Shanxi Inner Mongolia       Liaoning
             1              1              2              2              2              2
         Jilin   Heilongjiang       Shanghai        Jiangsu       Zhejiang          Anhui
             2              2              1              1              1              2
        Fujian        Jiangxi       Shandong          Henan          Hubei          Hunan
             1              2              2              2              2              2
     Guangdong        Guangxi         Hainan      Chongqing        Sichuan        Guizhou
             2              2              2              2              2              2
        Yunnan          Tibet        Shaanxi          Gansu        Qinghai        Ningxia
             2              2              2              2              2              2
      Xinjiang
             2
```

Figure 8: K-means&$C(g)$

Average silhouette method This method is based on the value of $s(i)$, whose formula is attached below:

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)}$$

the number of cluster is regarded to be the best when $s(i)$ approaches the maximum, the Figure 7 reveals that the maximum of $s(i)$ appears when the number of cluster is 2.

Clustergram method This is the method first proposed by Matthias Schonlau in his paper *The Clustergram: A graph for visualizing hierarchical and non-hierarchical cluster analyses*[3] published in The Stata Journal. Then Tal Galili[1] wrote the function specifically optimizes for K-means clustering. The Clustergram method gives a straight

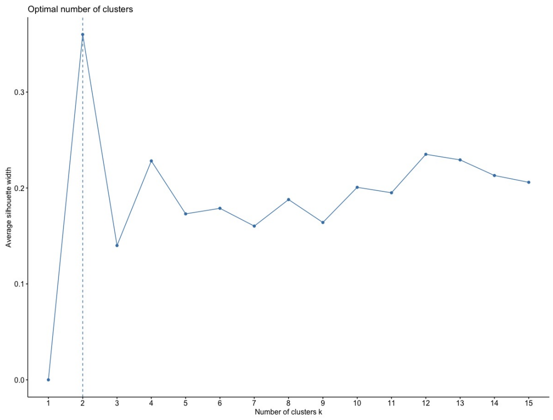


Figure 9: Average silhouette method

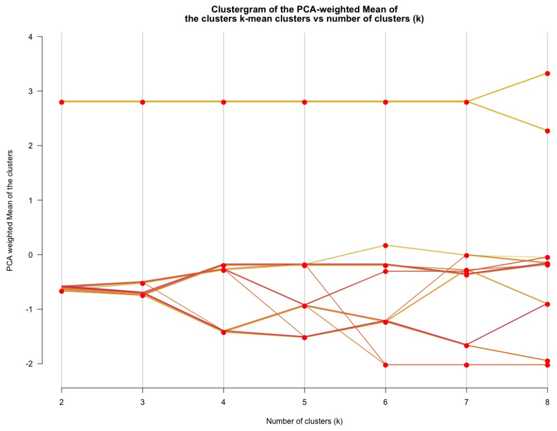


Figure 10: Clustergram of Kmeans

visualization of how each component goes when the number of clusters increases. Therefore, when the line between n clusters and $n+1$ clusters is more stable, the number n is more possible to be the optimal solution. Figure 9 demonstrates that the line is the most stable when the number of cluster is 2, however, we can observe that when $n=3$, although the line is not so stable compared to $n=2$, it still works much better than other number of clusters.

Determine the optimal number of clusters Based on the the criteria we discussed above, all the methods imply that $n=2$ is the best solution. This solution coincides with the solution derived by hierarchical clustering, but we do consider that if we apply 2 clusters, each cluster's size is relatively big, and also does not align with the current situation in China. Consequently, we visualize both $n=2$ and $n=3$ to see which one is more reasonable.

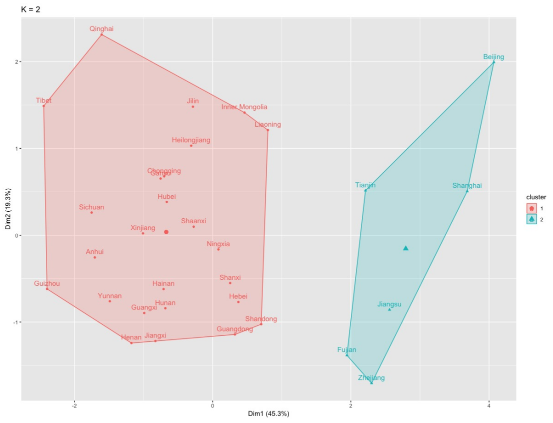


Figure 11: K-means with 2 clusters

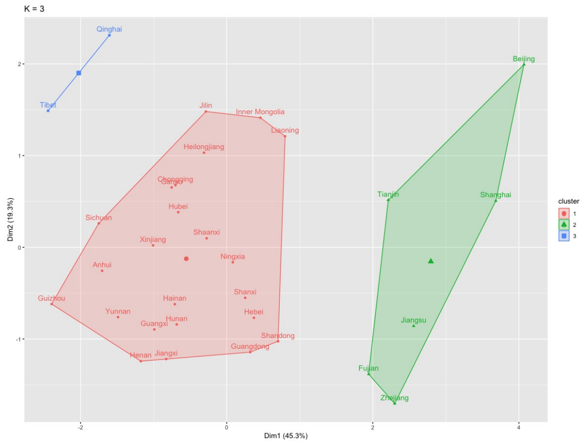


Figure 12: K-means with 3 clusters

When $n=2$, the outcome provided by K-means method is same as that in hierarchical clustering. When $n=3$, the difference reveals: K-means method separates **Tibet** and **Qinghai** whereas keeping **Ningxia** and **Chongqing** in their original group, and this satisfied the situation in China.

3.2.3 PAM&Mixed Model Clustering

To assure our outcome, we valid our model by PAM method and Mixed Model Clustering as well. In Mixed model clustering method, BIC value (Figure 13) shows that $n=2$ or $n=3$ should be the relatively optimal solution, but Figure 14 & Figure 15 both have some uncertainties that cannot be clustered. PAM method choose $n=3$ as the optimal solution, and it successfully categorized the first cluster, which contains those big cities such **Beijing** and **Shanghai**. Nonetheless, there still exists some overlap between the second cluster and the third cluster.

3.2.4 Summary and Comparison

In this part we applied three main clustering methods to our clustering process. Although the clustering result we get from PAM method is not ideal and have some overlap, we could still make comparison between any two of our clustering results. We use xtabs to do the comparison:

From the above result we can see that although the number of clusters in all the three methods is 3, the outcome of clustering is very different, meaning that the same province can be assigned in different clusters according to different methods.The method of complete linkage and K-means have 21 points in the same clustering result, while the

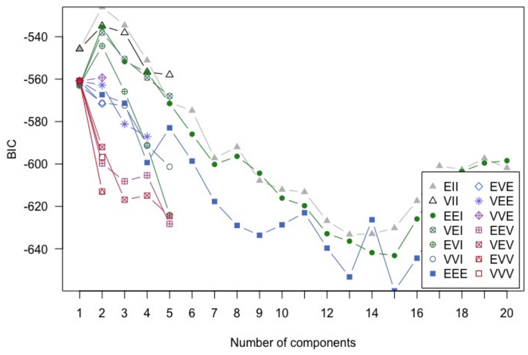


Figure 13: BIC Value

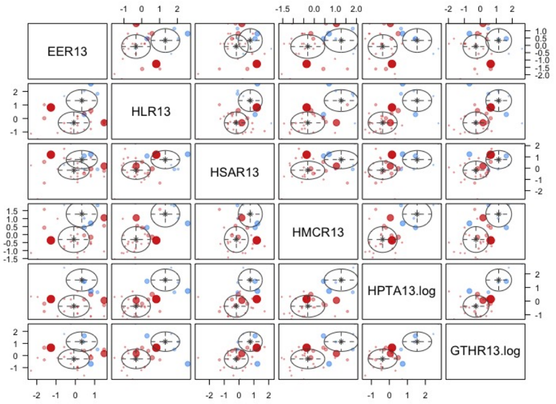


Figure 14: Uncertainties in Mclust when n=2

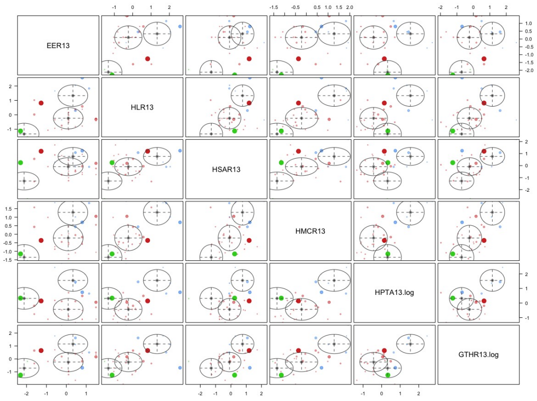


Figure 15: Uncertainties in Mclust when n=3

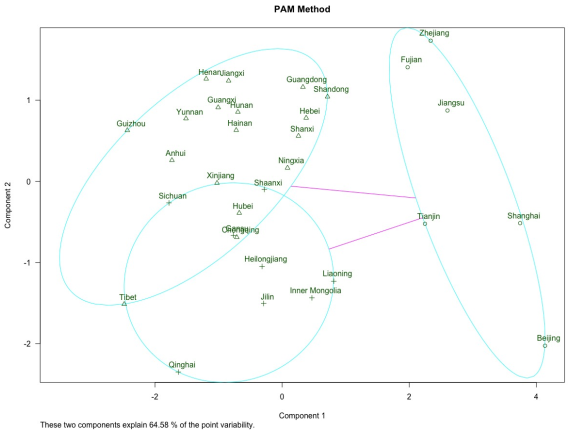
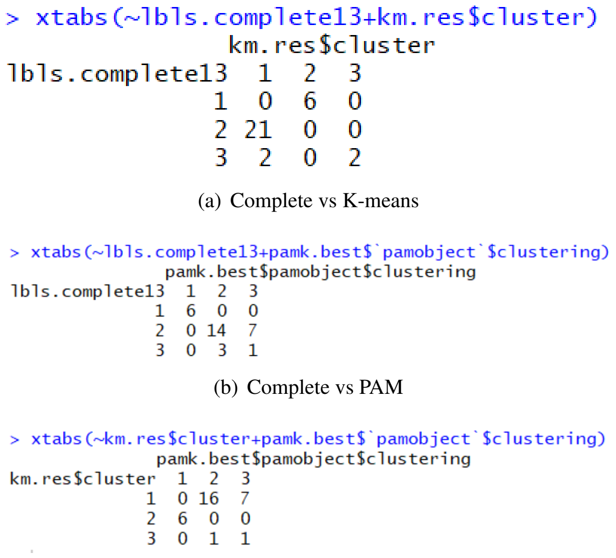


Figure 16: PAM Method



```
> xtabs(~lbls.complete13+km.res$cluster)
               km.res$cluster
lbls.complete13  1  2  3
              1  0  6  0
              2 21  0  0
              3  2  0  2
```

(a) Complete vs K-means

```
> xtabs(~lbls.complete13+pamk.best$`pamobject`$clustering)
               pamk.best$pamobject$clustering
lbls.complete13  1  2  3
              1  6  0  0
              2  0 14  7
              3  0  3  1
```

(b) Complete vs PAM

```
> xtabs(~km.res$cluster+pamk.best$`pamobject`$clustering)
              pamk.best$pamobject$clustering
km.res$cluster  1  2  3
             1  0 16  7
             2  6  0  0
             3  0  1  1
```

(c) K-means vs PAM

Figure 17: Comparison of three clustering results

method of K-means and PAM have 16 points in the same clustering result. All of these three methods can distinguish the first cluster, those highly developed cities or provinces: **Beijing**, **Shanghai**, **Jiangsu**, **Zhejiang**,**Tianjin**,**Fujian**. We also do notice that in both PAM clustering and hierarchical Clustering, **Chongqing** is categorized into the group where the education is not so developed, which is contrary to the current situation in China. Since **Chongqing** is one of the four Municipalities directly under the Central Government, it definitely owns the better education resources than other provinces. Based on the above comparison, we consider the clustering result of K-means as the optimal one in our analysis.

4 Discussion

4.1 Analysis on Additional Variables

After clustering, we mapped two additional variables which we have mentioned (Per capita GDP and Gini coefficient for education), as well as the clustering result of K-means.

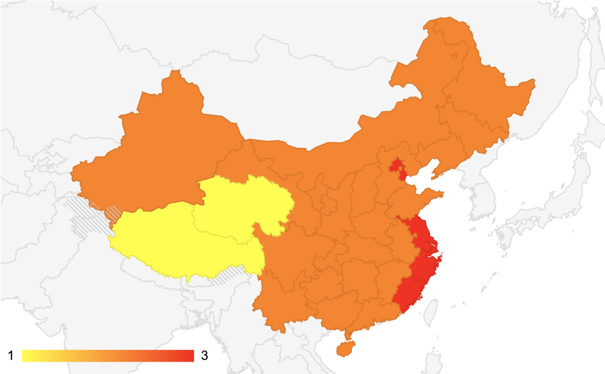


Figure 18: Map of China based on Clustering

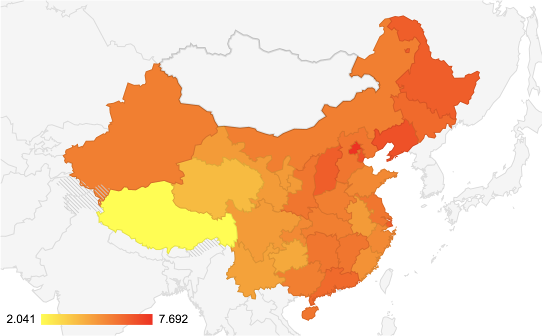


Figure 19: Map of China based on the reverse of Educational Coefficient

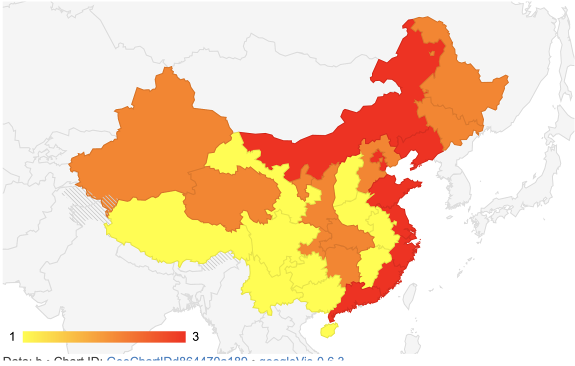


Figure 20: Map of China based on Per capita GDP

These three maps contain a lot of information:

- Figure 18: For a specific province, the clustering group result is closely related to the region in which it is located. The darker the color, the better the educational conditions in this province. The provinces in east region (6 provinces) usually have better educational conditions, while the provinces in the west region (especially **Tibet** and **Qinghai**) have poorer educational conditions. China's central region, northeast region and most western region have the same level of educational conditions (23 provinces).
- Figure 19: In order to correspond to the clustering result, we took the reciprocal of Gini coefficient, indicating that the higher the value, the darker the color and the fairer the education in that region would be. This map is somehow consistent with clustering results. A province having better educational conditions usually have higher education equality because the education resource are abundant so each student may receive more attention. But these two are not always positively correlated, we can see from the map that some southeastern provinces such as **Zhejiang** and **Fujian**, although with better educational conditions, have higher educational inequality that some northeastern provinces. This may suggest that there are other factors besides educational conditions that affect educational equality.
- Figure 20: The map of Per capita GDP does not exactly correspond to clustering results, which is a little bit surprising. Eastern provinces which enjoy a higher Per capita GDP always have better educational conditions, This is typical, indicating that there might be some positive correlation between these two variables. However, the high Per capita GDP does not guarantee high-quality educational conditions, meaning that educational conditions of a specific province is very complicated and cannot be predicted based on a single variable. We should make a combination of both economic and social factors.

4.2 Improvements

- Currently our project only explored the senior high school education, which may lead to unfairness to some certain extend. In the future we should include more variable related to the primary school or junior high school.

- In our model, the ”Quality Oriented Education” is mainly determined by the library area, the sports area as well as the number of multimedia classrooms. The criteria is based on our own experience rather than the strict research, so we should review more educational research paper to perfect our model.

- Due to the access restriction of the dataset, we cannot use the earlier data, it would be much more convinced if we can get the complete data and then compared the difference so that we can further explore the change that has been made in terms of the Chinese education.

References

[1] Tal Galili. Clustergram: visualization and diagnostics for cluster analysis (r code).

[2] National Bureau of Statistics of China. Chinese statistics yearbook 2013.

[3] Matthias Schonlau. The clustergram: A graph for visualizing hierarchical and non-hierarchical cluster analyses. *The Stata Journal,*, 3(3):316–327, 2002.

[4] the Ministry of Education of the PRC. Educational statistics in 2013.

[5] Xibo Fan Vinod Thomas, Yan Wang. Measuring education inequality: Gini coefficients of education. 2000.