Global Education Level Distribution and Prediction

Xiaofan Han, Zhicheng Ouyang, and Jian Guan

University of Michigan, Ann Arbor, MI

1 Introduction

Education is the most powerful instrument, not only for individuals to get employed and earn their lives but also for the society to overcome poverty, and promote equality and stability. In recent centuries, most countries and their governments have invested a lot to provide education resources to citizens, with tremendous progress achieved. The average years of schooling is a good measure to evaluate education attainment. People who stay in school longer are considered to hold more knowledge and ability to succeed in their life and are less likely to get involved in violence and crime. Large education years are therefore related to a relatively high education level.

The matrix data we used in our project is provided by Institute for Health Metrics and Evaluation (IHME)[1]. It covers the average years of schooling of people in different countries and regions over the age of 15 by sex and age group from 1970 to 2015. We are interested in how the education level varies with regions and age groups through this year as a reflection of efforts and outcomes of governments investing in education, and the global distribution, to which the politics and economics may contribute. Based on the results, we conducted some predictions to assess our model and forecast the developing trend in the future.

2 Methods

2.1 Regression

To and provide convincing evidence for our prediction, a regression model is needed. Firstly, we use the following binary quadratic regression model instead of a linear one for more accuracy, to estimate how mean years of education grow with time and age group:

$$y = p_0 + p_1 x_1 + p_2 x_2 + p_3 x_1^2 + p_4 x_1 x_2 + p_5 x_2^2 \tag{1}$$

In the equation, y is the mean years in school, while x_1 and x_1 represent years and age group id respectively. p_0 through p_5 are coefficients to identify. For each region, a set of these parameters is generated by the polynomial surface fitting function in Matlab. Predictions are based on the model.

2.2 Principal component analysis (PCA)

Principal component analysis (PCA) is a popular tool to reduce dimension by computing the principal component, especially for data visualization. In our regression model, each region corresponds to six parameters, which is a six-dimensional vector, making the PCA process necessary if we want to visualize the data in three-dimensional graphs. We also extract the first three principal components for clustering.

2.3 Cluster Analysis

Cluster analysis is used to classify observations into groups with high similarity within the cluster and large differences between each other clusters. It has a huge amount of applications in data analysis, machine learning, bioinformatics, image identification, and reconstruction. Hierarchical clustering is one of the simplest tools for clustering. Objects start with individual clusters and then merge together according to the distance until desired number of clusters are formed.

In our project, we apply hierarchical clustering to the three principal components obtained above. Then we plot the scattered points corresponding to each region in a 3D graph with respect to years, age groups, and the mean education years of people 45 years old in 1993, for comparing the absolute education level of regions, with various colors representing clusters. From the graph, the relationship and the trend we are interested in can be easily concluded.

2.4 Dynamic Mode Decomposition (DMD)

DMD is a data-driven and model-free algorithm extracting spatio-temporal patterns in the form of so-called DMD modes and DMD eigenvalues. The method relies simply on collecting snapshots of data x_k from a dynamical system at a number of times t_k, where $k = 1, 2, 3, ..., m$. We denote the sequence of snapshots collected by $\mathbf{X}$ and $\mathbf{X}'$where, $\mathbf{X}$ is the snapshots of system $\mathbf{X}$ is the time-shifted snapshot matrix of $\mathbf{X}$. Let us suppose that there is an unknown linear operator $\mathbf{A}$ such that $\mathbf{X}' = \mathbf{AX}$.

To obtain operator $\mathbf{A}$, compute SVD of matrix $\mathbf{X} = \mathbf{U\Sigma V}^*$, one can write $\mathbf{X}' = \mathbf{AU\Sigma V}^*$. Then we Define $\tilde{\mathbf{A}} = \mathbf{U}^*\mathbf{AU} = \mathbf{U}^*\mathbf{X}'\mathbf{V\Sigma}^{-1}$ Compute the eigendecomposition of $\tilde{\mathbf{A}}$. Let $\mathbf{W}$ is the matrix of eigenvectors, and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. Each eigenvalue λ_i is a DMD eigenvalue. 4. Compute the DMD modes,

$$\mathbf{\Phi} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{W} \tag{2}$$

Each column of $\mathbf{\Phi}$ is a DMD mode $\mathbf{\Phi}$ corresponding to eigenvalue λ_i. Then one can approximate the system by

$$\mathbf{x}(t) \approx \mathbf{\Phi}\exp(\mathbf{\Omega}t)\mathbf{b} \tag{3}$$

where $\mathbf{\Omega} = \mathrm{diag}(\omega)$,is a diagonal matrix whose entries are the eigenvalues $\omega_k = \ln\lambda_k/\Delta t$, Δt is the sampling interval between $\mathbf{X}$ and $\mathbf{X}'$. $\mathbf{b} = \mathbf{\Phi}^{-1}\mathbf{x_1}$ is the initial condition.

3 Results

3.1 Factor Analysis

Education level are relevant with four parameters according to the data: the age, sex, country, and the year. In order to figure out the most influential parameters, we performed the factor analysis.

Two factor is chosen to indicate the dominant parameters. We used the "factoran" function in the MATLAB.factoran computes the maximum likelihood estimate (MLE) of the factor loadings matrix Λ in the factor analysis model.

$$\Lambda = \begin{bmatrix} -0.0016 & 0.9975 & 0.0006 & -0.3549 \\ 0.4153 & 0.0038 & -0.1374 & 0.9322 \end{bmatrix}^T \tag{4}$$

The (i, j) entry indicates the dependency of i-th parameter on j-th factor. We could concluded that the 2nd the 4th parameters is highly relevant with the 1st and 2nd factor, indicating that they are the most influential parameters.

3.2 Regression in years of schooling of males of different countries and age groups

By surface fitting and PCA, three principal components were extracted, followed by hierarchical clustering. Regions were classified into three, four, five, and six groups respectively, to figure out the best clustering performance. The results are visualized in Figure 1. When there are three or four groups, regions in the bottom with low or high developing rates were not separated clearly, especially the green ones in the first plot (k=3) and the yellow ones in the second plot (k=4). As for six groups in the fourth plot (k=6), there is not much difference between the green ones and the red ones. Compared to others, the plot with five clusters gives a better view of how the overall education performance varies from region to region without redundant groups. Therefore five classes are chosen for the subsequent analysis.

We plotted the clustering results in a 3D graph with respect to years, age groups, and median situations,

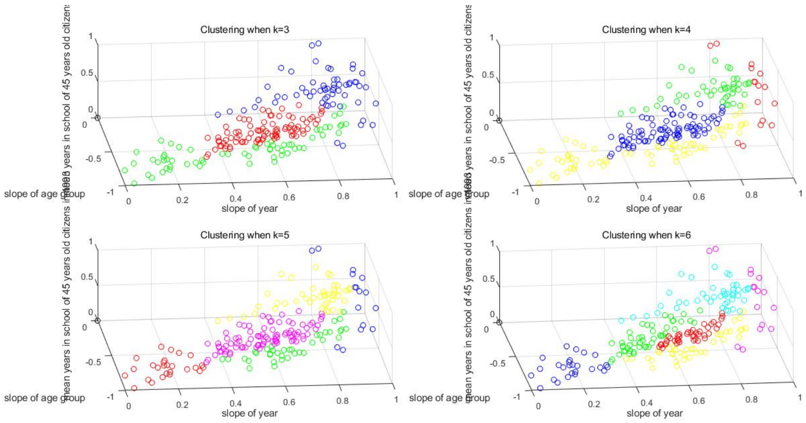

Fig. 1: **Comparing of different number of clusters.** Different colors indicate different clusters.

as shown in Figure 2, in which each color represents a group. As the difference in schooling years between age groups simply decreases with the increase of education level, we paid more attention to the education effort through years.

The blue points correspond to the regions with the most rapid progress over the past 50 years. Yellow, violet, and green ones all have relatively high increasing rates, with different fundamentals. Red ones represent the group with low education levels and developing progress.

Those with a higher absolute value and lower divergence of mean years in school between age groups mainly appearing in blue and yellow are mostly in Europe, which is the origin of modern science, creating a social environment friendly to learning and thinking for citizens. Many countries in central Asia are also in these two categories, maybe due to the scientific and cultural exchange with Europe. A similar situation happens to high-income North Latin America, especially in the US and Canada, since they have spent a lot of their GDP on education and research, based on the resources and intelligence brought by European immigration several hundreds of years ago.

Regions in violent and green contain most countries and regions from south Latin America, Oceania, and

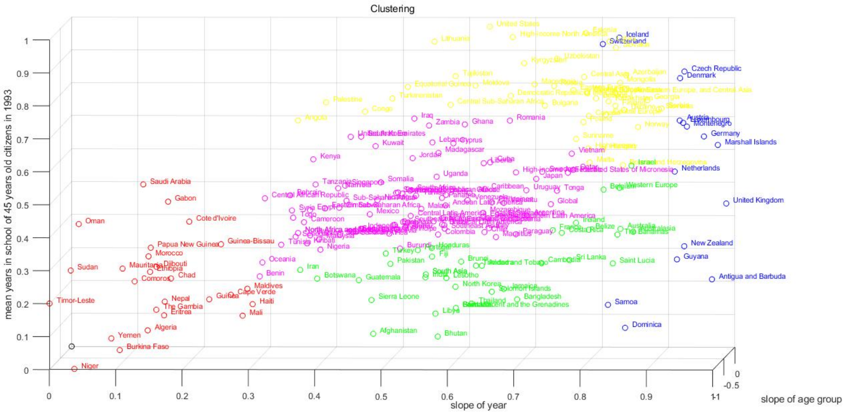

Fig. 2: **Mean years of education vary with years.** Blue: Fast developing regions. Yellow: High education level regions. Violet: Sub-high education level regions. Green: Medium education level regions. Red: Low education level regions.

the rest of Asia, north Latin America, as well as half of Africa. These regions hold a desirable developing rate mainly due to compulsory education acts of their governments in these years. As GDP keeps increasing around the globe, the education expense of countries and regions grows quickly, providing more chances for people to equip themselves with knowledge. A peaceful and stable society also contributes a lot.

Regions from the rest of Africa are the main part of the group in red. Some of these countries are still suffering from conflicts, poverty, and lack of natural resources. The large discrepancy in schooling years between different age groups indicates an imbalance in their education. As a whole, people around the world still have a long way to go to eliminate inequality and imbalance worldwide.

3.3 DMD Prediction

Based on the clustering result, we pick out five representative counties: United Kingdom, Canada, China, India, and Sudan. Then we made prediction based on the data of these country.

DMD analysis is performed in this part. The system has only conjugate 2 eigenvalues, whose values are $1.0042 \pm 0.0155i$. This result is reasonable because the education Using Eq.(2)(3) to generate DMD approximate. As figure 4 shows, the prediction perfectly match the real situation provided enough data. The predicted curves also consolidates the conclusion we made above. Education levels in Britain are not only high in absolute terms, but also growing rapidly. Though Canada had the highest education level at 1970s, but it grew slower than other countries during 2000s. One could find that in our prediction, Canada's education level would be surpassed by Britain in absolute value. China has a mediate education level and rapid growth. India and Sudan had low education rate in 1970s, but India had a faster growth, hence it surpassed Sudan and is clustered into the medium education level region.

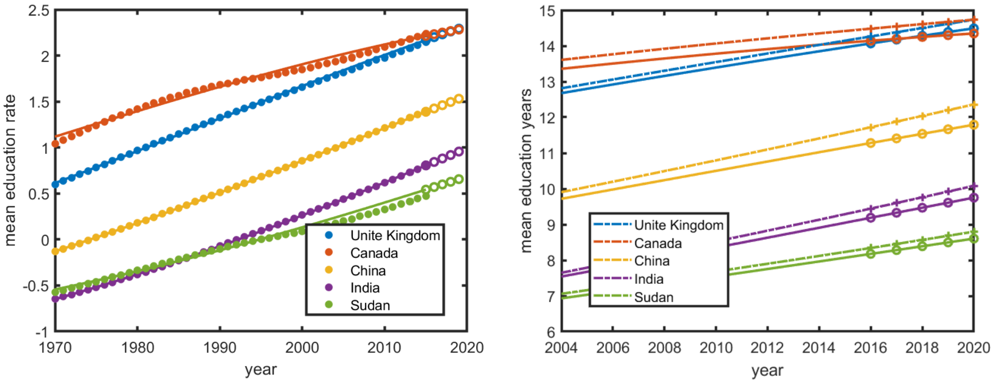

Fig. 3: **A comparison of DMD, regression predicted result and real data**. The solid dots in the left panel are the real data (standardized value), while the curves are generated by DMD. In the right panel, we zoom in the last decade and provide five year prediction of education level represented by circles and plus sign. The solid line is the result generated by DMD while the dashed line is based on regression

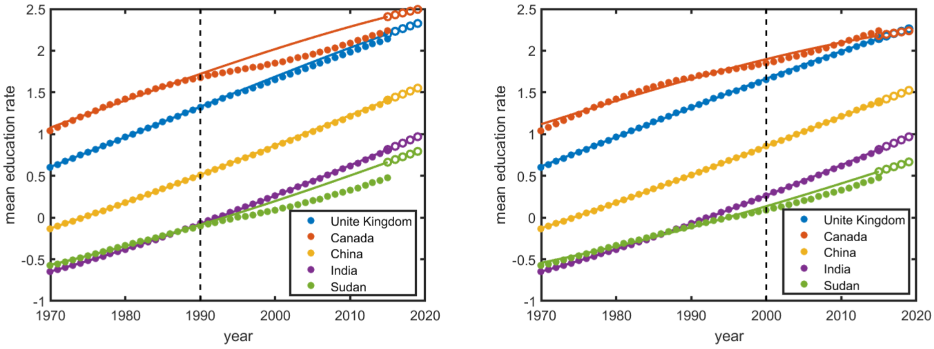

Fig. 4: **DMD prediction.** The vertical dashed line indicate how many data had been used. In the left panel, the prediction was made base on data from 1970 to 1990. In the right panel, data from 1970 to 2000 were used

To eliminate the education level difference arise from different ages, we also used the "Age Standardized Education Years Per Capita" data provided by IHME [1] to do quadratic regression and provide five year prediction with the regression model. The results are shown in Figure 5. Generally, we find that United Kingdom (Fast developing regions) and Canada (High education level regions) have the highest aged standardized education level. China (sub-high education level regions) has a moderate aged standardized education level. India (Medium education level regions) and Sudan (Low education level regions) have the lowest aged standardized education level, but their education level increase the fast. Besides, we also calculate the male-female age standardized education level ratio to indicate the education difference between genders in different regions. We find that initially the regions which have high education level (e.g. United Kingdom and Canada) have the lowest education difference between genders, the regions which have low education level (e.g. Sudan and India) have extremely high education differences between genders, but the gender difference in these regions also decrease the fastest with time. Generally, the gender

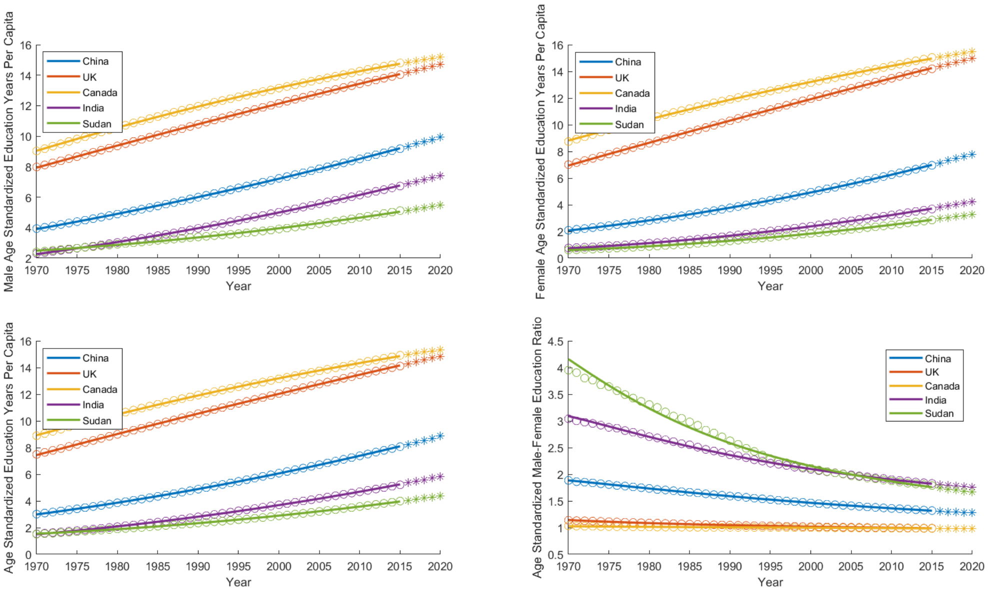

Fig. 5: **Age standardized education years per capita for the five representative counties.** The circles are real data points from 1970 to 2015. The solid lines are the fitting curves. The stars are the five year prediction of aged standardized education level. Up-left: Male age standardized education level. Up-right: Female age standardized education level. Bottom-left: Age standardized education level considering both genders. Bottom-right: Male-female age standardized education level ratio.

4 Conclusion

In this project, we analyzed the global progress and distribution in education. First, we established a binary quadratic regression model to examine how education years varied through years and age groups. We applied PCA and hierarchical clustering to divide these regions into five groups and studied the reasons behind. Next, we used the data from five selected countries and did prediction based on DMD, regression algorithm. The result showed that DMD could make precise prediction for education level for the next 15 years using data of previous 30 years. Besides, we also find that the education difference between genders has a reverse trend with education level among different regions.

References

1. "Global Educational Attainment 1970-2015." GHDx, 1 Jan. 1970, https://ghdx.healthdata.org/record/ihme-data/global-educational-attainment-1970-2015.