# Regional Inequalities in Education in China: Could Virtual Learning Changes the Situation? *

Yuming Liu†

June 2020

**Abstract**

Because of the unbalanced developments between coastal and inland provinces and between urban and rural areas, regional imbalance has been a major issue in China for decades since the reform and opening-up. Spatial inequalities are in areas like education and health care. Our paper would focus on the regional inequalities in education. We built several models using data collected from the National Yearbooks and applying Machine Learning algorithms to explain the causes of education inequalities. From the results, we determined whether the popularization of virtual learning could help reduce inequalities and reshape the educational resource allocation in the next decade.

*keywords:* China, Inequality, Education, Schooling, Virtual Learning, Statistic, Spatial Analysis, Machine Learning, Prediction.

*JEL classification:*

*Thank my family, my friends, and Dr. Richard Evans for offering me remarkable supports during my anxious period!

†The University of Chicago, MACSS, (773) 329-7830, yumingl@uchicago.edu.

# 1 Introduction

Put introduction here. You'll probably want some references here like Auerbach (1996) or DeBacker et al. (2015). These citations use the `natbib` package above and require a call to the `bibliography` command just before the appendix section.

# 2 Theory

# 3 Data

## 3.1 Data Overview and Sources

The data we collected could be separated into five parts: National Survey data of Population aged 6 and above by Sex, Educational, Attainment and Region from 2012 to 2018, National data of Numbers of Schools by Level and Region from 2000 to 2018, National data of Disposable Personal Income (DPI) by Region in 2018, National data of Consumption by Field and Region in 2018, and the "2018 China Online Education Industry White Paper" from iiMedia Research. The Chinese National Data could be accessed online from the official website of the National Bureau of Statistics, and it could also be viewed on the China Statistics Yearbooks. The white paper from iiMedia Research could be accessed from their official website as well. Since the data from 2019 is lack of some important information in educational categories, the data from 2018 is more comprehensive for the research.

## 3.2 Data Review

Some papers from the literature review section also used the data in same fields. For instance, Fleisher, etc. (2010) [4] used the National Survey data of Education Attainment to measure the difference of population attending college or above between the interior provinces and the coastal provinces. Yang, etc. (2014) [6] also used this data, and they got the income and fund data from the China Statistic Year Book.

## 3.3 Data Cleaning

The raw data collected did not include the Education Gini Index, Average Year of Schooling and other variables. Therefore, we calculate the Education Gini Index and Average Year of Schooling from the raw Education Attainment data, which only had the total size of surveyed population and the size of population attained each level of education. We divided each population of the groups with the total surveyed population for all provinces and used the percentage to generate the Education Gini Index and Average Year of Schooling.

We calculate the Average Year of Schooling from the function:

$$AYS = \sum_{i=1}^{n} y_i p_i$$

where

$AYS$ is the Average Year of Schooling based on education attainment distribution;

$p_i$ stands for the proportion of population with a certain level of schooling;

$y_i$ is years of schooling of a specific level;

$n$ is the number of levels.[4]

On the other hand, we generate the Education Gini Index based on the function:

$$E_L = \frac{1}{\mu} \sum_{i=2}^{n} \sum_{j=1}^{i-1} p_i \mid y_i - y_j \mid p_j$$

where

$E_L$ is the education Gini based on education attainment distribution;

$\mu$ is average years of schooling for the sample population;

$p_i$ and $p_j$ stand for the proportions of population with certain levels of schooling;

$y_i$ and $y_j$ are years of schooling of specific levels;

$n$ is the number of levels.[1]

The Virtual Learning data we selected from the "2018 China Online Education Industry White Paper" includes three categories: population of online learning users,

percentage of online learning users by ago group in 2018, Attention rate of virtual learning by region in 2018.

# 4 Methods

## 4.1 Spatial Patterns of Key Variables

Firstly, we must understand the causes of regional inequalities in educations. The significant variables indicating education inequalities are the Average Year of Schooling and the Education Gini Index. To show the spatial patterns of the variables, the most common used method is to generate the quantile maps with the geographic information systems (GIS).

**Figure 1: Average Year of Schooling (AYS) by Region in China, 2018**

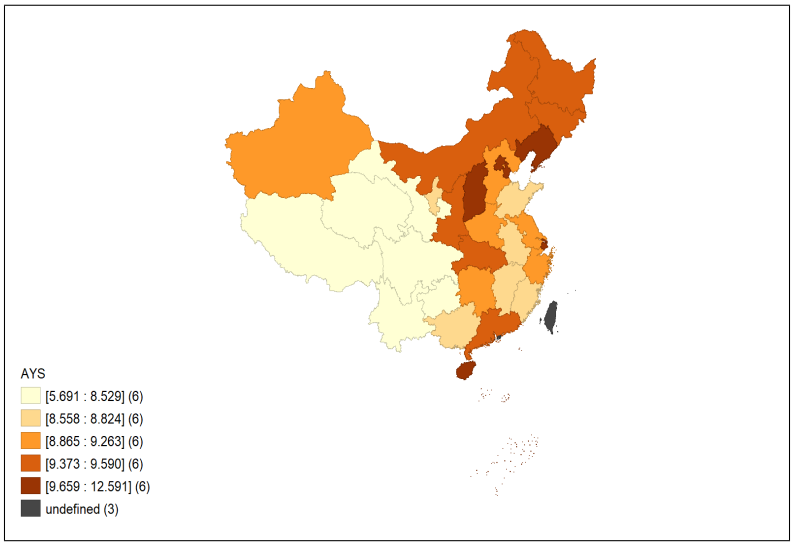
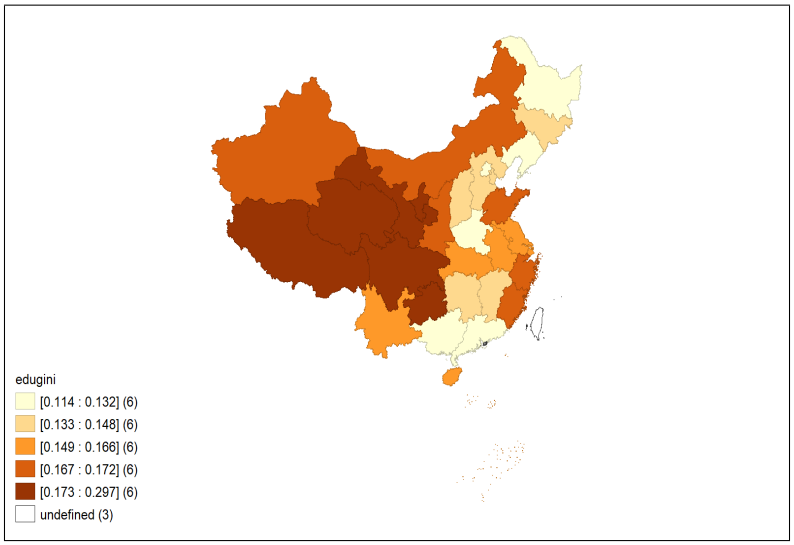



Figure 1 and Figure 2 show the quantile maps with five bins, and each bin contains six regions. From the maps, we observe that the areas with the lowest AYS and the areas with the highest Education Gini share an outstanding overlap. Interior provinces in the west are those that suffered from the most extreme education inequalities. However, the interesting fact is that not all coastal provinces have higher AYS, and

**Figure 2: Education Gini Index by Region in China, 2018**




some of them even have the higher Education Gini than some interior provinces. The general observations emphasize that though there are clearly some remarkable spatial patterns for education inequalities, the spatial correlation of these two variables could only explain a limited part of the causes. Besides AYS and Education Gini Index, we have to include other variables to the analysis.

## 4.2 Measuring the Relationship

We filtered the data and selected six other variables with relatively high correlation with the AYS and the Education Gini. Two variables to mention are the number of primary schools and the number of universities. We determine to combine them as '5/2' to indicate the ratio of universities to primary schools. We expect the ratio to reflect the competitiveness for students to attend colleges or above. For the rest, 'rural_pct' is the percentage of rural population; 'edu_con' is the amount of education and entertainment consumption per person; 'income' is the DPI; 'virt_att' is the media attention of virtual learning.

From Table1, almost all variables have at least weak correlation with the others. It means that running multi-variable regression would not influence the significant of independent variables too much. Since income and '5/2' didn't show the apparent

correlation with Education Gini, we would not include these variables for the regression model that set Education Gini as the dependent variable and consider whether add them or not based on further significance tests.

**Table 1: Pearson correlation Among Selected Variables**

| | AYS | edugini | income | edu_con | rural_pct | 5/2 | virt_att |
|---|---|---|---|---|---|---|---|
| AYS | 1.00 | -0.66 | 0.75 | 0.83 | -0.88 | 0.78 | 0.53 |
| edugini | | 1.00 | -0.26 | -0.44 | 0.48 | -0.24 | -0.32 |
| income | | | 1.00 | 0.92 | -0.88 | 0.86 | 0.75 |
| edu_con | | | | 1.00 | -0.91 | 0.83 | 0.63 |
| rural_pct | | | | | 1.00 | -0.82 | -0.58 |
| 5/2 | | | | | | 1.00 | 0.56 |
| virt_att | | | | | | | 1.00 |

## 4.3 Model Construction

We would build three model systems. The first two would be applied to independently predict the values of the AYS or the Education Gini Index, and the last one is a combination of two regressions with no common independent variable. It would predict the AYS and use the results to then predict the Education Gini. For instance, we perform the regression model of the AYS as

$$AYS_i = \beta_0 rural\_pct_i + virt\_pct + \epsilon$$

where $i$ is the region, *virt_pct* is the national ratio of virtual learning users to population, which functions as a constant. Then we perform the regression model of the Education Gini as

$$Edugini_i = \beta_0 + \beta_1 AYS_i + \beta_2 edu\_con_i + \beta_3 virt\_att_i + \epsilon$$

where $AYS_i$ is the previous equation.

The beginning step would always be using linear regression to test the significance of variables. Then we would record the significant variables and employ those to other

Machine Learning algorithms to compare the scores of the models. Then the last step would be using the top three models with highest scores to generate the prediction results for each system.

# 5 Results

## 5.1 Regression Models and Results on AYS

For the 31-region OLS model, the strongest predictor is Education Gini, and the other strong predictors are shown in Table2.

**Table 2: OLS Model Resutls on AYS**

| | Value | $SE$ | $t$ | $p$ |
|---|---|---|---|---|
| virt_pct (constant) | 78.8833 | 3.623 | 21.776 | 0.000 |
| rural_pct | -0.0400 | 0.012 | -3.206 | 0.003 |
| 5/2 | 17.9967 | 5.988 | 3.006 | 0.006 |
| edugini | -12.5880 | 2.490 | -5.056 | 0.000 |
| Multiple $R$-squared | 0.893 | | | |
| Model $p$-value | 0.00 | | | |

We dropped the DPI and the amount of education and entertainment consumption per person for the OLS regression model since these variables are not significant. The national ratio of virtual learning users to population performs as a constant in this model since the data is not provincial. From Table2, as we expect according to Table1, less academic competitiveness would result in higher '5/2' statistics and higher AYS. Moreover, higher ratio of virtual learning users to population would also help to increase the AYS. This means that the popularization of virtual learning and the sharing of more virtual learning resources could help regions with lower AYS to narrow their gaps with other regions, and offering people opportunities to attend online college-level or professional courses could improve the accessibility to education for those regions.

Table 3 demonstrates a comparison between the scores and predictions of a list of models using different Machine Learning algorithms. For the predictions, we used

the predicted ration of virtual learning users per population in 2020 and controlled other variables.

**Table 3: Model Scores and Predictions on AYS**

| | 2018 Value | Random Forest | Nerual Network | SVR |
|---|---|---|---|---|
| Score | - | 0.9362 | 0.8987 | 0.6451 |
| Henan | 8.865 | 8.833 | 8.879 | 8.655 |
| Tibet | 5.691 | 7.152 | 6.711 | 8.439 |

## 5.2 Regression Models and Results on Education Gini

Similarly, we have that the strongest predictor is the AYS, and the other strong predictors could be found in Table 4.

**Table 4: OLS Model Resutls on Education Gini Index**

| | Value | $SE$ | $t$ | $p$ |
|---|---|---|---|---|
| Intercept | 0.4037 | 0.036 | 11.225 | 0.000 |
| AYS | -0.0270 | 0.005 | -5.348 | 0.000 |
| Education&Entertainment Consumption | -2.474e-05 | 1.28e-05 | -1.934 | 0.065 |
| DPI | 2.922e-06 | 9.08e-07 | 3.219 | 0.004 |
| Number of Primary Schools | -3.425e-06 | 8.72e-07 | -3.927 | 0.001 |
| Meida Attention of Virtural Leanring | -0.1754 | 0.127 | -1.385 | 0.178 |
| Multiple $R$-squared | 0.793 | | | |
| Model $p$-value | 0.00 | | | |

Although the DPI don't have a strong correlation with Education Gini, we found it is a strong predictor. Education and Entertainment Consumption per person includes the consumption in both categories. Therefore, containing entertainment consumption could decrease the significance, but since the p-value is still between 0.05 and 0.10, we determined to keep this variable. The reason why adding the Media attention of Virtual Learning and Number of Primary Schools is that besides others selected, those are the two variables with the greatest significance. We also tried adding the ratio of number of schools per population. However, it wasn't a successful attempt.

**Table 5: Model Scores on Education Gini Index**

| | Random Forest | Neural Network |
|---|---|---|
| Score | 0.8593 | 0.8987 |

In addition, Table 5 emphasizes a comparison between the scores of a list of models using different Machine Learning algorithms.

## 5.3 Combined Regression Models and Results

We have the Combined OLS models shown as below Table 6.

**Table 6: Combined OLS Model Resutls**

| | Value | SE | t | p |
|---|---|---|---|---|
| virt_pct (constant) | 80.5174 | 2.227 | 36.153 | 0.000 |
| rural_pct | -0.0854 | 0.008 | -10.234 | 0.000 |
| Multiple R-squared | 0.783 | | | |
| Model p-value | 0.00 | | | |
| Intercept | 0.4446 | 0.066 | 6.729 | 0.000 |
| AYS(Model) | -0.0331 | 0.010 | -3.424 | 0.002 |
| DPI | 4.655e-06 | 1.13e-06 | 4.132 | 0.000 |
| Education&Entertainment Consumption | -3.769e-05 | 1.48e-05 | -2.545 | 0.017 |
| Number of Primary Schools | -3.487e-06 | 1.06e-06 | -3.303 | 0.003 |
| Media Attention of Virtual Learning | -0.2975 | 0.153 | -1.939 | 0.064 |
| Multiple R-squared | 0.698 | | | |
| Model p-value | 0.00 | | | |

**Table 7: Combined Model Scores and Predictions**

| | 2018 Value | Random Forest | Neural Network |
|---|---|---|---|
| Score | | 0.8773 | 0.8233 |
| Beijing | 0.129 | 0.1435 | 0.029 |
| Henan | 0.132 | 0.1439 | 0.020 |
| Tibet | 0.297 | 0.2503 | 0.2247 |

The results shows that the AYS generated from the percentage of rural population and the ratio of virtual learning users to population is sightly less significant, but the significance of other variables such as Education and Entertainment Consumption per person and Media Attention of Virtual Learning is remarkably improved. Table 7

emphasizes a comparison between the scores and predictions of a list of models using different Machine Learning algorithms.

# 6 Conclusion

# References

**Auerbach, Alan J.**, "Dynamic Revenue Estimation," *Journal of Economic Perspectives*, Winter 1996, *10* (1).

**DeBacker, Jason, Richard W. Evans, Kerk L. Phillips, and Shanthi Ramnath**, "Estimating the Hourly Earnings Processes of Top Earners," Technical Report, Mimeo 2015.

# APPENDIX

## A-1 Some Appendix

You can put appendices here at the end of the paper using section commands.