# Could Virtual Learning Changes our Next Generation? Regional Inequalities in Education in China *

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:* C53, I24, I28.

*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

In the past 30 years, China has become one of the fastest developing countries around the world since the beginning of reform and opening-up. With the booming economy, education in China also keeps improving as the government made nine-year compulsory education universal nationwide. However, similar to the unbalanced economic development, education also faces remarkable inequalities. The disparities in access to education between rural and urban areas and between coastal and inland provinces are the major cause of educational inequalities in China.

In 2014, China finished building the universal 4G network. It means that people could all have fast internet to watch live videos and communicate from video-meeting platforms. From the same year, virtual education has blossomed and attracted its first group of users. In the beginning, the services were expensive and only available for a limited number of users. Moreover, most people, who were able to deal with the high cost, refused to pay a thousand dollars a year to take online courses that could also be taught in physical classrooms.

Nowadays, with the technological supports from the 5G network and artificial intelligence, virtual learning becomes popular and affordable for most people using smartphones. Virtual learning now has the advantages of being flexible and convenient. It is rich in resources and combined with Virtual Reality (VR) and Augmented Reality (AR) to meet personalized needs and solve difficulties to access physical education.

This paper aims to study the presentation of regional inequalities in education in China and analyze the causes. Further, we apply spatial analysis to find the spatial patterns of education inequalities. Then we would examine whether the popularization of virtual learning could help to fix regional inequalities in education.

## 2 Background and Literature Review

The basic methods of the Gini index and the Gini Coefficients of Education were from the article of Thomas et al. (1999). The article introduced the original Gini index function and then later applied the modified function to determine education inequality. The Gini Coefficients of Education have variables including levels of schooling, the proportions of the population with certain levels of schooling, and the years of schooling at different educational attainment levels. The paper employed the education Gini index to measure inequality in educational attainment. Also, the paper presented two methods, direct and indirect, of calculating the Education Gini index. It generated a dataset on Education Gini for population age over fifteen, for 85 countries from 1960 to 1990. The paper took spatial and temporal approaches to the data, and they found that education inequality for most of the countries has been declining during the three decades. The analysis also showed that globally different regions still had significant variances of the Education Gini index. Finally, after controlling for the initial income level, the growth of GDP per capita (PPP) is negatively related to inequality in education and positively related to the average years of education of the labor force.

Thomas's analysis provided us a clear guideline of using educational data to formulate the Educational Gini index, and it also provided us some inspirations of using social data such as gender and religion. Moreover, although the analysis didn't use many economic variables, it still highlighted the significance of income and PPP. Geographically, Thomas, etc. only focused on national data. Another paper from Zhang et al. (2015) concentrated on the educational inequalities between rural and urban areas. The paper applied data from the China Family Panel Survey (CFPS) and the Rural-Urban Migration in China (RUMiC) survey to compare the education performance of rural children, children of rural-to-urban migrants, and urban children between 2009 and 2010. The results they got showed that the education performance of rural children and migrants' children is remarkably lower than which of their urban counterparts. The paper mainly suggested that differences in personal

attributes such as nutrition and parenting style would have significant effects on the performance of the children. Therefore, even while the paper provided a regression analysis to explain the educational inequalities between urban and rural areas, the variables for the regression are selected from a psychological perspective. The paper is particularly insightful for policymakers to reduce the educational inequalities between the areas, and the data they collected was also the inspiration for our research since it almost covered all facts about regional education and the reflections from the students.

Unlike the paper from Zhang aiming at both urban and rural areas, De Brauw and Giles (2008) was trying to show that people migrating from the rural areas to the urban areas were more likely to have their children attending higher-level education. It shows that people from the rural areas usually did not choose to attend high schools or colleges because they hoped to use their physical abilities and the chances to work in the urban areas to earn more wages for their family instead of relying on incomes from agricultural production. Another fact was that the possibilities for rural students to attend a good university in China were remarkably low. Therefore, learning knowledge from higher-level education seemed less attractive for most of the students living in rural areas. Moreover, since urban kids were more likely to attend higher-level schools, their families would be more willing to spend money investing education resources, and for rural families, they would just send their kids to work in urban areas to have jobs instead of continuing supporting them to go to high school or colleges. On the other hand, incomes for rural families are also addressed here as an important variable. Since the college tuitions were extremely expensive for some families and the student loans system was not well-established, rural families could fall into poverty when supporting their children attending colleges. Hence, most people migrating from the rural areas to the urban areas would enthusiastically support their children to fulfill their abilities to go to outstanding colleges when these parents didn't have their opportunities to do so.

Regional inequalities in economic growth have been an extraordinary issue in China. Fleisher et al. (2010) analyzed how regional growth patterns in China depend

on the differences in human capital, infrastructure capital, and foreign direct investment. The variances of real per-capita GDP between the wealthiest coastal provinces and the poorest interior provinces are remarkable, and the ratio of the real per-capita GDP between them was 8.65. Fleisher found that even though the proportion of adults who had at least some senior high school education or above was not shown an outstanding variability for most regions, the proportion of the individuals with at least college degrees in the costal and northeast regions was much higher than the others. The paper stated that educational inequalities are an important part of regional inequality. Since China's economic growth didn't benefit its provinces and regions equally, the paper suggested that the high degree of regional income inequalities and the high degree of education inequalities have an interactional relation. The results of the paper inspire our research to concentrate on variables that could represent the phenomenon of regional income inequalities and other inequalities of economic growth.

Zhang and Kanbur (2005) were the trailblazers who applied systematically spatial analysis to present the facts on inequalities in education and health care. The paper also addressed that social inequalities had increased "substantially since the reforms and opening-up began. Yang et al. (2014) applied the Education Gini index from Thomas et al. (1999) to demonstrate the scales of education inequalities in China. They calculated it for every four years from 1996 to 2008 for all provinces using data from the China General Social Survey. The table of the Education Gini coefficients clearly showed which provinces had the values below the national one and which provinces had higher education levels than the national average level. Our research intends to use the same data that Yang's paper used but for different years from 2018 till now. The method of showing the table of Education Gini coefficients would also be referred to by our research. Moreover, besides the general spatial differences for the 28 provinces, the paper also included the variables such as gender, hukou (rural or urban), and income, to indicate the potential factors for education inequalities in the nation. Some social facts the paper brought were that rural areas had lower Education Gini coefficient and lower average years of schooling (AYS) than urban, Female had

higher Education Gini coefficient but similar AYS than male, groups of people older than 26 in 2006 usually had lower AYS but higher Education Gini coefficients than groups younger than 26. The paper calculated the educational investment as the sum of government appropriation for education, funds from private schools, donations and fund-raising for running schools, income from teaching research and other auxiliary activity, and other educational funds. They got the data from the China Statistic Year Book. For our research, I would like to more specifically focus on only the average educational investments per household, and the government appropriation for education. The paper worked on a semi-log regression model based on OLS to analyze which factors contributed more to education inequalities. They also used the Shapley decomposition to determine that. Based on the results, they found that hukou and income contributed to education inequalities most. It provides us an insight into highlighting these factors in our research as well. Yang's paper offered us a detailed guideline and a solid installation of the regression model between the Education Gini Index and other economic variables, and from this paper, we would add more variables and machine learning methods to find a more accurate relation between these factors.

Xiao and Liu (2014) drew extraordinary maps to show the inequalities between the two provinces in China. I would take a similar GIS approach for visualizing the data for our research and provide maps to cluster the hotspots of education resources in different areas.

The above references would help us to construct our models and analysis of education inequalities in China. On the other hand, the paper from Van Raaij and Schepers (2008) and the paper from Tang and Wang (2017) stated how companies and schools were applying virtual learning to products. Van Raajj published the paper in 2008 when people in China were not adapted to developing computer technologies in education and showed the anxiety people had when learning from an unfamiliar system. Tang's paper more directly aimed at how virtual learning could create more educational resources and help people, especially migrant workers, to satisfy their demands of educating themselves on their electronic devices. These two papers would

help us to develop a hypothesis on whether popularizing virtual learning could reduce education inequalities and how.

# 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 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 et al. (2010) 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 et al. (2014) 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.(Fleisher et al., 2010)

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;

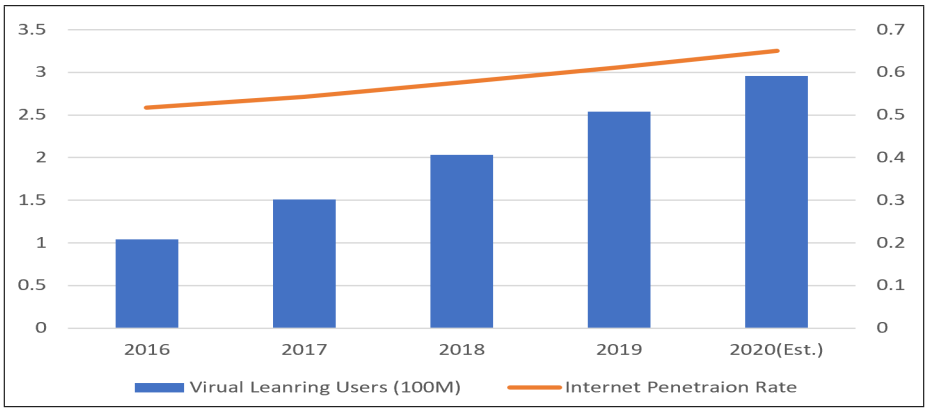
$p_i$ and $p_j$ stand for the proportions of population with certain levels of schooling;

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

$n$ is the number of levels.(Thomas et al., 1999)

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. Figure 1 shows the trends of development of virtual learning users and the internet penetration rate.

**Figure 1: The development of Internet Penetration Rate and Virtual Learning Users in China (2016-2020)**



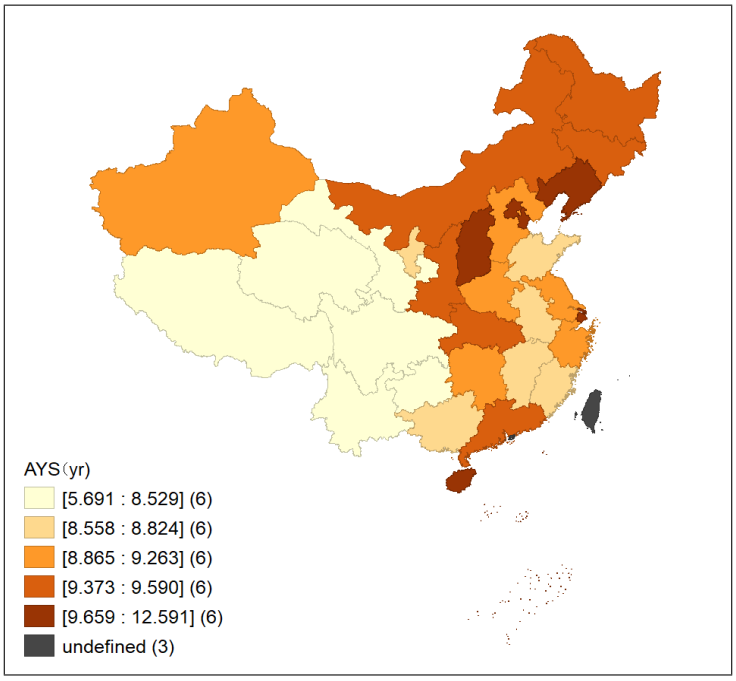

# 4 Methods

## 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 2 and Figure 3 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 some of them even have 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.

**Figure 2: Average Year of Schooling (AYS) by Region in China, 2018**




## 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 Table 1, 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, at first, not include these variables for the regression model, which sets Education Gini as the dependent variable, and consider whether add them or not based on further significance tests.

**Figure 3: Education Gini Index by Region in China, 2018**

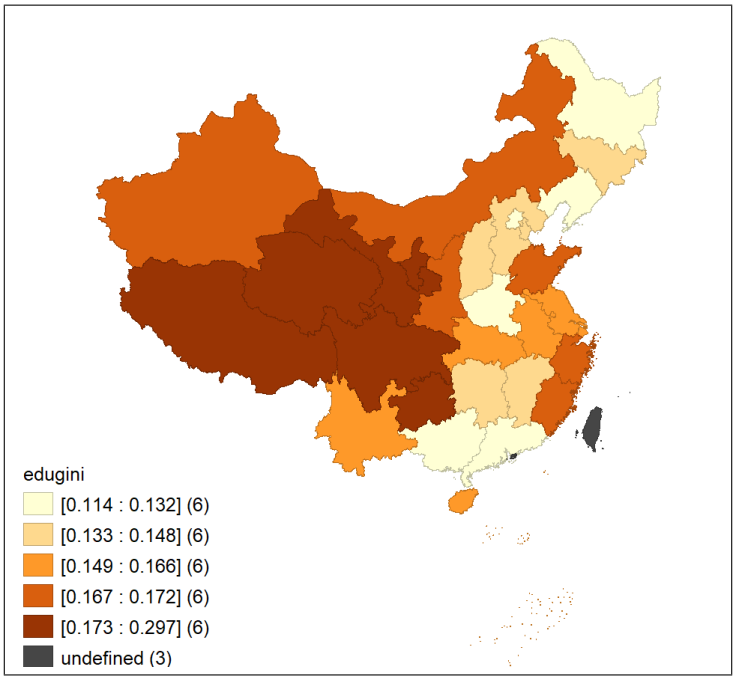



**Table 1: Pearson Correlation Among Selected Variables**

| | 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

$$AYS_i = \beta_1 rural\_pct_i + virt\_pct + \epsilon$$

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. We control the values of all other variables besides the ratio of virtual learning users to population.Then we apply the estimated ratio in 2020 to the Machine Learning models and get the prediction results.

# 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 Table 2.

**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 Table 2, as we expect according to Table 1, 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 ratio 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.

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

**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 | | | |

ratio of number of schools per population. However, it wasn't a successful attempt. In addition, Table 5 emphasizes a comparison between the scores of a list of models using different Machine Learning algorithms.

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

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

## 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.

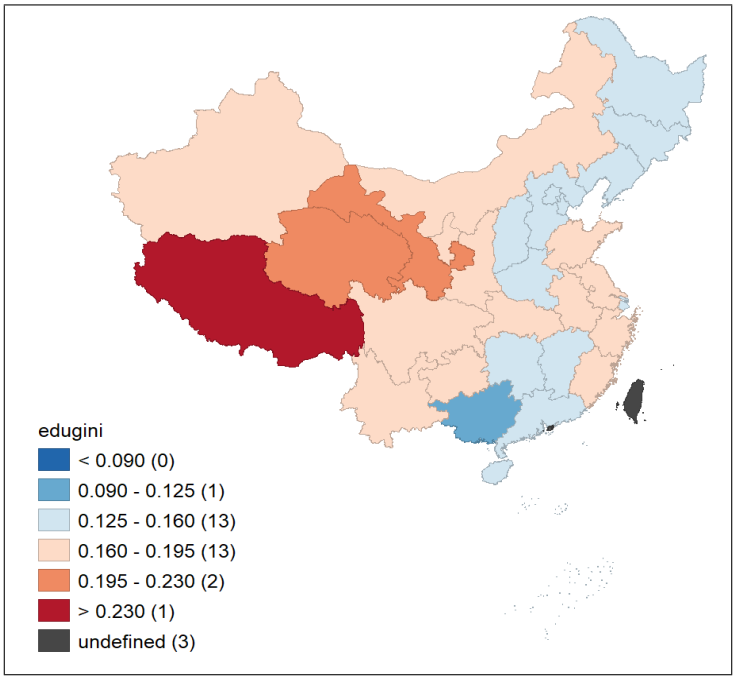
# 6 Conclusion

From the prediction results of the Machine Learning Algorithms, we determine that the increase of virtual learning resources and virtual learning users would help to eliminate the regional inequalities in education. Although the western provinces remained their Education Gini Indexes higher than the rest of China, the gaps between them could be narrowed by the popularization of virtual learning. The results from the first model imply that the AYS for the provinces with higher Education Gini would have an outstanding improvement with more virtual learning users while which of the provinces with lower Education Gini could remain the same.

The SD maps shown by Figure 4 and Figure 5 demonstrate that increasing the number of virtual learning resources and virtual learning users would balance the nationwide education and decrease the Education Gini for all areas that had their values higher than the mean in 2018. Twenty-two out of thirty provinces now have similar Education Gini, and their values are all in a healthy range. On the other hand, since education is always competitive and self-motivated, Thomas et al. (1999) stated that no region can have an Education Gini Index approximate to 0 when people have freedom. Therefore, for those provinces with lower Education Gini in 2018, the

prediction results show that increasing virtual learning users would increase their Education Gini Indexes. This means that people willing to use virtual learning would pursue further education, and that just would stretch their distance with the others in years of education.

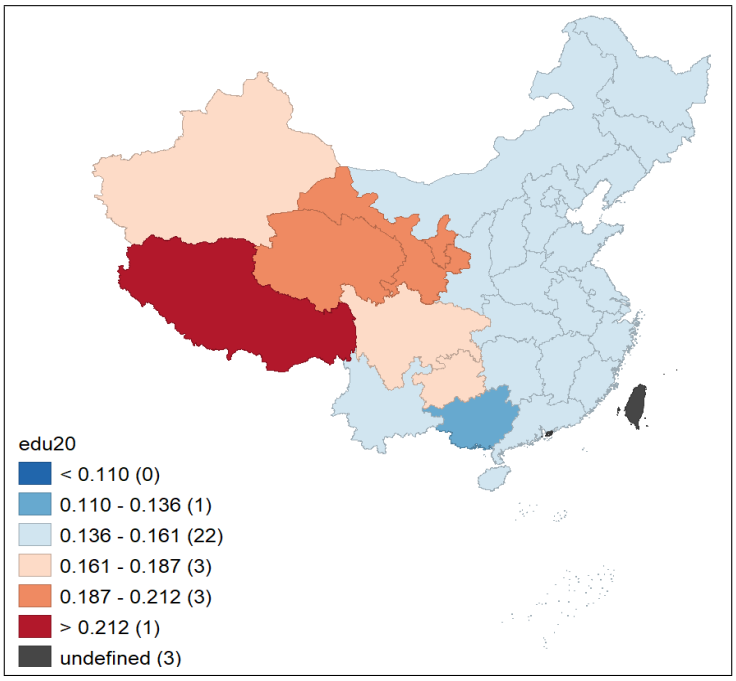
**Figure 4: Standard Deviation of Education Gini Index by Region in China, 2018**




# 7 Discussion

The paper only focused on the effects of virtual learning to regional inequalities in education. Although increasing virtual learning has been proven by the results as a way to fix inequality, other variables such as rural percentage and number of universities are also significant to AYS or Education Gini. However, those variables, especially the rural percentage by region, are extremely difficult to estimate since the causality is complicated and related to the changes in policies. Based on the data we have, the best way to construct the model would be only using the most regional data and apply the data from 2019 to test the accuracy of the model when it

**Figure 5: Predicted Standard Deviation of Education Gini Index by Region in China, 2020**




becomes available. Moreover, the 2020 General Census data in China is going to more comprehensive for our model testing. Using those could provide us more conclusions.

Since the data of virtual learning was still limited because of User Privacy Terms when writing the paper, the lack of provincial data caused us difficulties to further develop the spatial model. During the COVID-19 pandemic, nearly all schools in China switched to virtual learning for a period. Virtual learning is acknowledged by the Chinese government, and platforms such as Zoom and Dingding had an enormous increase of users. In those months, people's experience could change their opinions on this field. Many schools and organizations are working on surveys to ask people's attitude towards virtual learning. Therefore, at the end of 2020, we believe more data in this area would be published. It could allow us to adjust the variables and construct more accurate models.

# References


**Brauw, Alan De and John Giles**, *Migrant opportunity and the educational attainment of youth in rural China*, The World Bank, 2008.

**Fleisher, Belton, Haizheng Li, and Min Qiang Zhao**, "Human capital, economic growth, and regional inequality in China," *Journal of development economics*, 2010, *92* (2), 215–231.

**Raaij, Erik M Van and Jeroen JL Schepers**, "The acceptance and use of a virtual learning environment in China," *Computers & Education*, 2008, *50* (3), 838–852.

**Tang, Yan'er and Simin Wang**, "The Construction of Virtual Learning Community Based on the New Generation of Migrant Workers' Needs for Continuous Education," *Modern Distance Education Research*, 2017, (3), 12.

**Thomas, Vinod, Yan Wang, and Xibo Fan**, *Measuring education inequality: Gini coefficients of education*, The World Bank, 1999.

**Xiao, Jin and Zeyun Liu**, "Inequalities in the financing of compulsory education in China: A comparative study of Gansu and Jiangsu Provinces with spatial analysis," *International Journal of Educational Development*, 2014, *39*, 250–263.

**Yang, Jun, Xiao Huang, and Xin Liu**, "An analysis of education inequality in China," *International Journal of Educational Development*, 2014, *37*, 2–10.

**Zhang, Dandan, Xin Li, and Jinjun Xue**, "Education inequality between rural and urban areas of the People's Republic of China, migrants' children education, and some implications," *Asian Development Review*, 2015, *32* (1), 196–224.

**Zhang, Xiaobo and Ravi Kanbur**, "Spatial inequality in education and health care in China," *China economic review*, 2005, *16* (2), 189–204.