ANALYSIS OF TUITION GROWTH RATES BASED ON CLUSTERING AND REGRESSION MODELS

Long Cheng and Chenyu You

Department of Electrical, Computer and Systems Engineering
Department of Mathematical Sciences
Rensselaer Polytechnic Institute, Troy, NY, USA

ABSTRACT

Tuition plays a significant role in determining whether a student could afford higher education, which is one of the major driving forces for country development and social prosperity. So it is necessary to fully understand what factors might affect the tuition and how they affect it. However, many existing studies on the tuition growth rate either lack sufficient real data and proper quantitative models to support their conclusions, or are limited to focus on only a few factors that might affect the tuition growth rate, failing to make a comprehensive analysis. In this paper, we explore a wide variety of factors that might affect the tuition growth rate by use of large amounts of authentic data and different quantitative methods such as clustering and regression models.

KEYWORDS

Tuition Growth Rate, K-means Clustering, Linear Regression, Decision Tree, Random Effect

1. INTRODUCTION

Education is undoubtedly helpful for countries' development [1], especially considering nowadays many countries have transferred from a labour-intensive economy to a knowledge-based economy [2]. Research study in [3] shows that higher education is becoming increasingly expensive while the return on its investment is falling, which will eventually adversely affect how people view the importance of education. Meanwhile, the rising tuition growth rate plagues tens of thousands of families every year [4], which motivates us in this paper to explore what are the key factors that cause such high tuition growth rate.

A fixed-effects model using cross-sectional and time-series date is proposed in [5] to explore how different racial groups react to the tuition. A method using multivariate regression analysis to investigate the relationship between enrollment rates and tuition is studied in [6] with the consideration of different races, enrolment rates and states. [7] studies how the tuition growth rate affects personal decision on higher education and shows that even though both higher student scholarship and lower tuition could increase the students enrolment rate, the actual amount of money that students have to pay by themselves is a more important factor that affects their decision. [8] applies the investment theory and the consumption approach to study the demand for higher education with the consideration of the tuition fee. And the result shows that the

DOI : 10.5121/ijdkp.2016.6401

demand for higher education react positively to family income increase and negatively to tuition fee increase. The effect of socioeconomic status of students on education opportunity and tuition fee is explored in [9]. The conclusion in [10] indicates the tuition fee has less impact than some other micro-econometric factors on college enrollment. It is also showed that the change of the tuition fee might have different impacts on different geographic locations, which are also affected by the tax rate [11].

Though some preliminary research on tuition fee analysis has been done in above papers, they fail to present a very detailed analysis using large amounts of raw data and considering a wide variety of factors. This paper takes United States as an example and explores many different factors, including both in-school ones and macro-economic ones, that might have impacts on tuition levels. And we find out that the possible in-school factors that determine tuition price might be university ranking, class size, percentage of full-time faculty, college acceptance rate, classroom equipment, financial need, and the number of full-time students. In the macro-economic prospect, unemployment rate, price level, government policy and the average income level might make great effects on the rise of college tuition in the United States. Other possible factors that might contribute to the increment of tuition rates include geography and population.

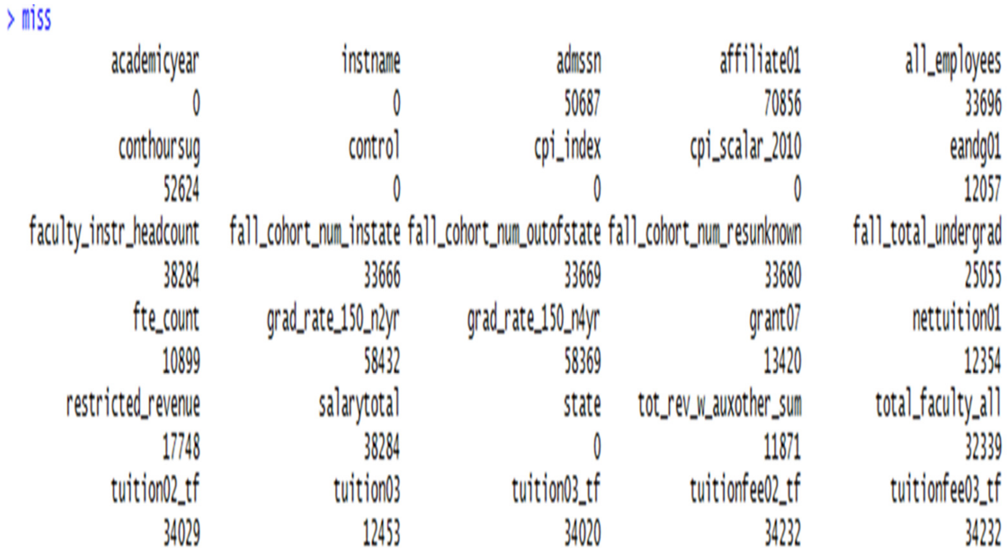
2. DATA DESCRIPTION

A reliable dataset from National Center for Education Statistics IPEDS Analytics: Delta Cost Project Database [12] is used in this paper. This is a longitudinal database derived from IPEDS finance, enrollment, staffing, completions and student aid data for academic years 2000-2001 through 2009-10.

There are 950 variables used in the dataset to record information about tuition fee, state name, school name, employee number, and so on. All the variables are carefully reviewed and finally 30 variables are selected in this paper to perform the analysis. For example, considering the fact that many graduate students are part-time students, whose tuition fees vary a lot even though they are in the same graduate program, so we decide to delete all data related with graduate students, and only focus on undergraduate students. Meanwhile, schools are divided into three categories: private school, public school with in-state tuitions, and public school with out-of-state tuitions, for the convenience of analysis. After the initial variable selection procedure, those variables are again inspected to see if they contain too many missing values. There are five variables (see Figure 1) that contain more than 40000 missing values, which are more than half of the total number (the total number is around 80000). Even though there are some existing methods to add the missing data, the accuracy of the analysis could be impaired if we use those methods to add so many missing data for those five variables. So we decide to delete those five variables and only use the other 25 variables. This paper will further explain how these variables are selected and processed in details in Part 3 ANALYSIS.

3. ANALYSIS

This data set contains around 1000 schools' information. Since most students only care more about schools they are familiar with, our paper only focuses on those well-known universities. So we set up the minimum bounds of 25% quantile and 75% quantile SAT scores as 400 and 600, for both Math and Verbal sections. After this SAT score selection, there are 370 schools left.



```
> miss
           academicyear                instname                  admssn             affiliate01           all_employees 
                      0                       0                   50687                   70856                   33696 
            conthoursug                 control               cpi_index         cpi_scalar_2010                 eandg01 
                  52624                       0                       0                       0                   12057 
faculty_instr_headcount fall_cohort_num_instate fall_cohort_num_outofstate fall_cohort_num_resunknown  fall_total_undergrad 
                  38284                   33666                   33669                   33680                   25055 
              fte_count       grad_rate_150_n2yr      grad_rate_150_n4yr                 grant07            nettuition01 
                  10899                   58432                   58369                   13420                   12354 
     restricted_revenue             salarytotal                   state    tot_rev_w_auxother_sum       total_faculty_all 
                  17748                   38284                       0                   11871                   32339 
           tuition02_tf               tuition03            tuition03_tf         tuitionfee02_tf         tuitionfee03_tf 
                  34029                   12453                   34020                   34232                   34232 
```

Figure 1. Missing values of the selected 30 variables

In addition, 9-year data is used in this paper to analyze the trend of tuition rates. For schools with complete 9-year tuition information, we calculate the annual tuition growth rate using [(this year / the previous year)-1] for each year. For schools with only 8-year tuition information, we predict the annual tuition growth rate by [the square root of (the latter year/ the previous year)-1] for the gap year. For instance, if one school misses 2008 tuition information, we predict 2008 tuition growth rate by [square root of (tuition 2009/tuition 2007) -1]. Those schools with less than 8-year tuition data are deleted from the dataset. After above steps, the dataset now contains 9-year tuition information for about 300 universities in 42 states. Lastly, the employee/student ratio and faculty /student ratio are calculated using the raw data and added into the final dataset as two new variables.

For a clear presentation, tuition growth rates in the 42 states will be shown using figures according to three categories (private schools, public schools with in-state tuitions, and public schools with out-of-state tuitions). In these figures, x-axis stands for the year, and y-axis stands for the tuition growth rate. For each state, the 25% quantile, 50% quantile, and 75% quantile of the tuition growth rates of all the schools belonging to the same category are calculated and shown in the following figures.

(1) Private schools

Figure 2 shows the tuition growth rates from 2002 to 2009 for private schools in 42 states. And we remove graphs of some states that don’t have significant changes of tuition growth rates. In the end, there are 16 graphs left, which means that there are 16 states that have apparent trend changes in tuition growth rates.

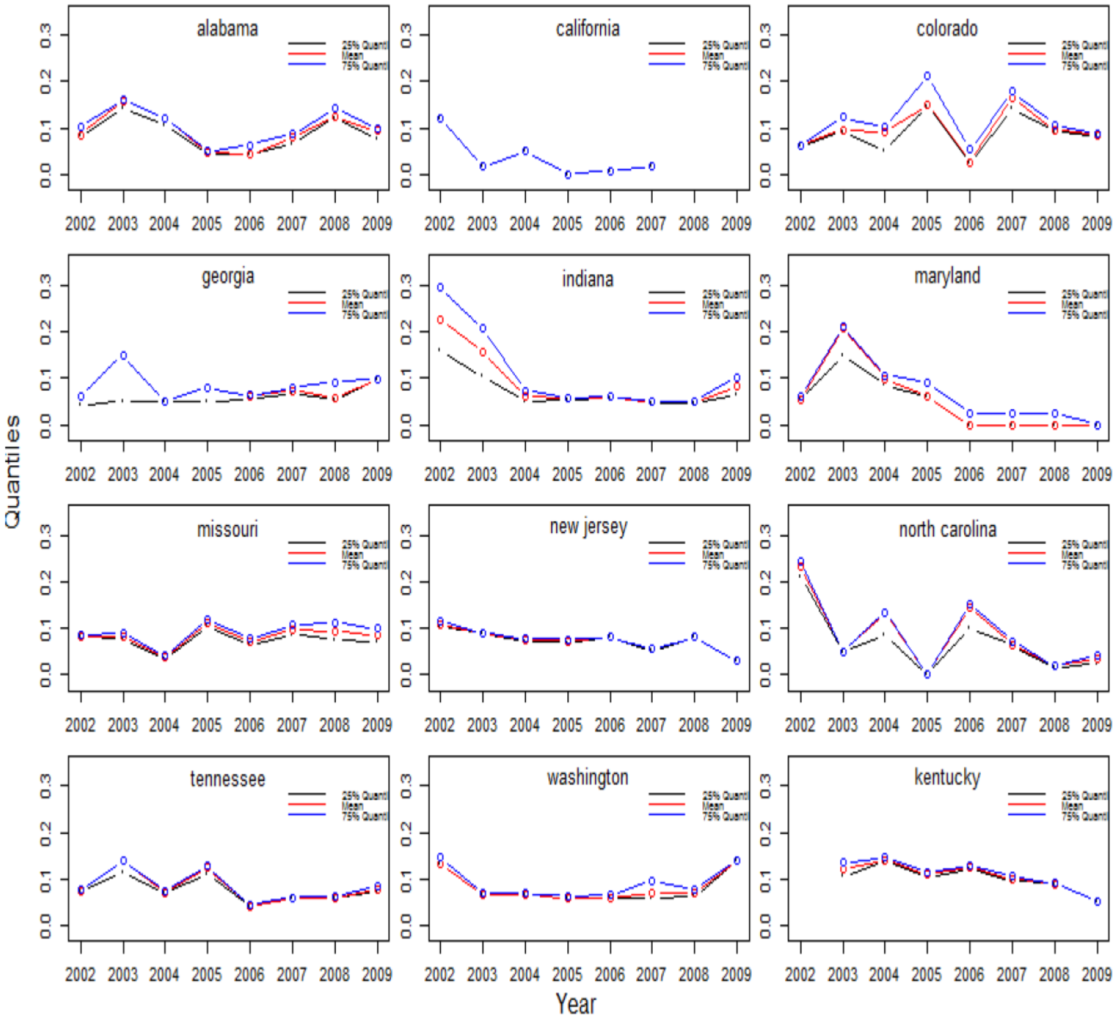


Figure 3. Tuition growth rates from 2002 to 2009 for public schools with in-state tuition in 12 states

Since all public schools belong to the UC system in California and all those schools have the same tuition growth rate, there is only one line in the graph for California.

To figure out the reason why Alabama, Colorado, Indiana, Maryland, and North Carolina have significant changes of tuition growth rates, the original dataset is checked and it is noticed that only a few schools belongs to this category in these five states. So changes of tuition growth rates in one school affect the results of the whole state a lot.

(3) Public schools with out-of-state tuition

Similarly, there are 14 states that have apparent trend changes in tuition growth rates, as shown in Figure 4. We find out that almost all these 14 states, except for Texas, have slow changes during those years.

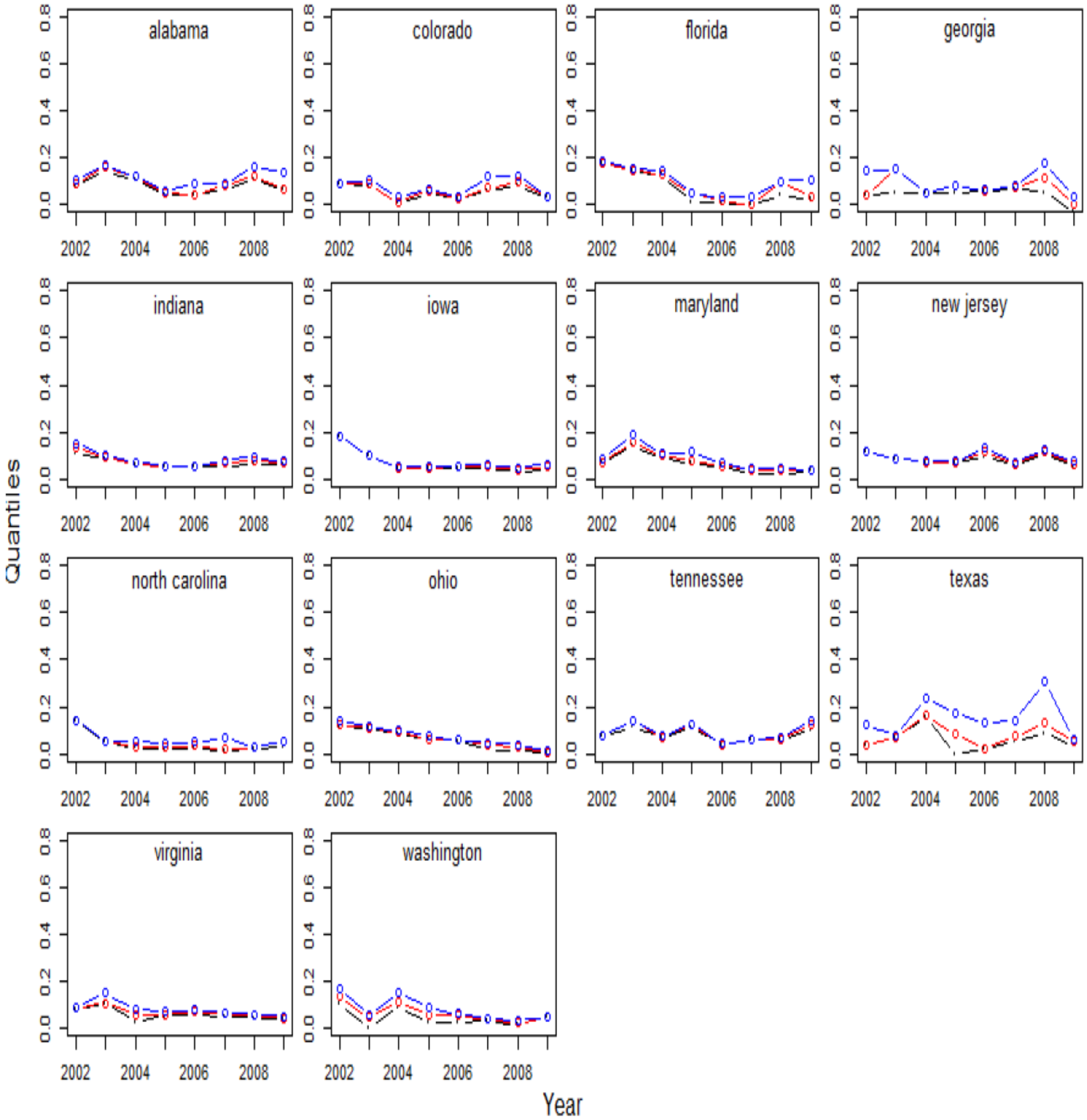


Figure 4. Tuition growth rates from 2002 to 2009 for public schools without-state tuition in 14 states

4. Model Development

4.1 Clustering Analysis

K-means clustering is utilized in this paper to analyze the dataset based on four criteria: quantiles of all years, quantile in every single year, tuition growth rates, and tuition fees. For each of those criteria, different numbers of clusters are chosen. We also draw state maps to show the clustering results for each category for all years. And in all these maps, the smaller cluster number represents the lower tuition growth rate, and the larger cluster number represents the higher tuition growth rate. For instance, Cluster #1 represents for the lowest tuition growth rate.

4.1.1 Using Quantiles of All Years

(1) Choose the number of clusters

From Figure 5 and Figure 6, we decide to use 4 clusters for private schools, and 6 clusters for public schools with out-of-state tuition. There is no need to cluster public schools with in-state tuition, since the corresponding figure is only a horizontal line.

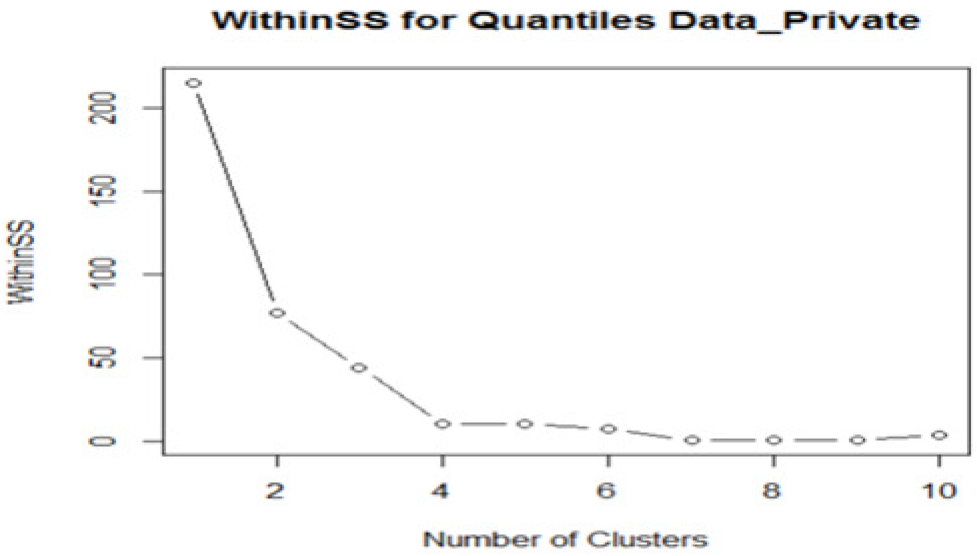


Figure 5. WithinSS for quantiles data (private schools)

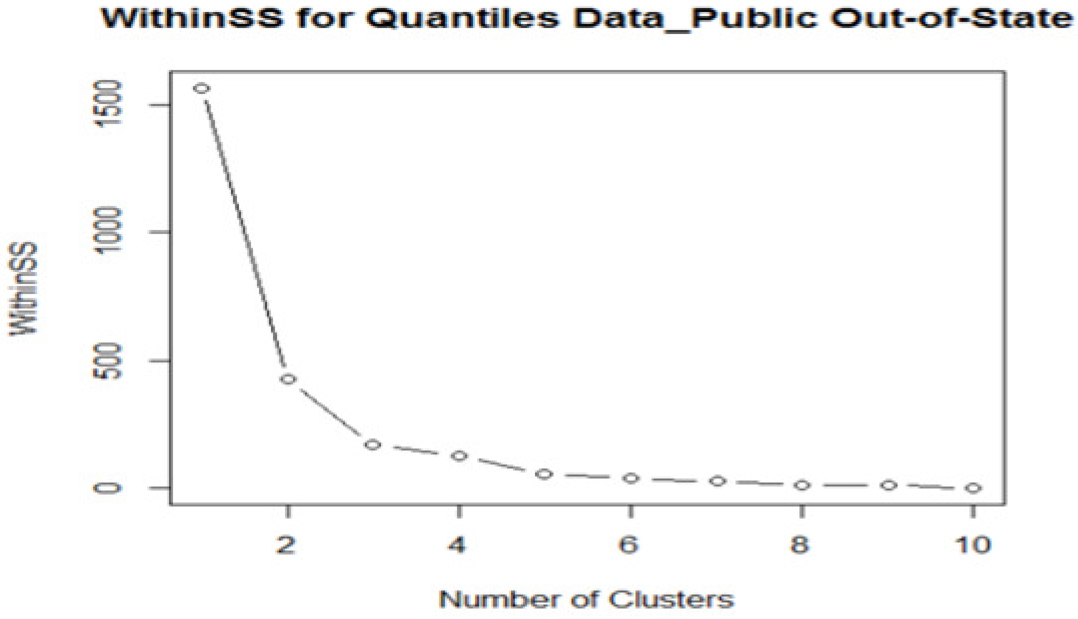
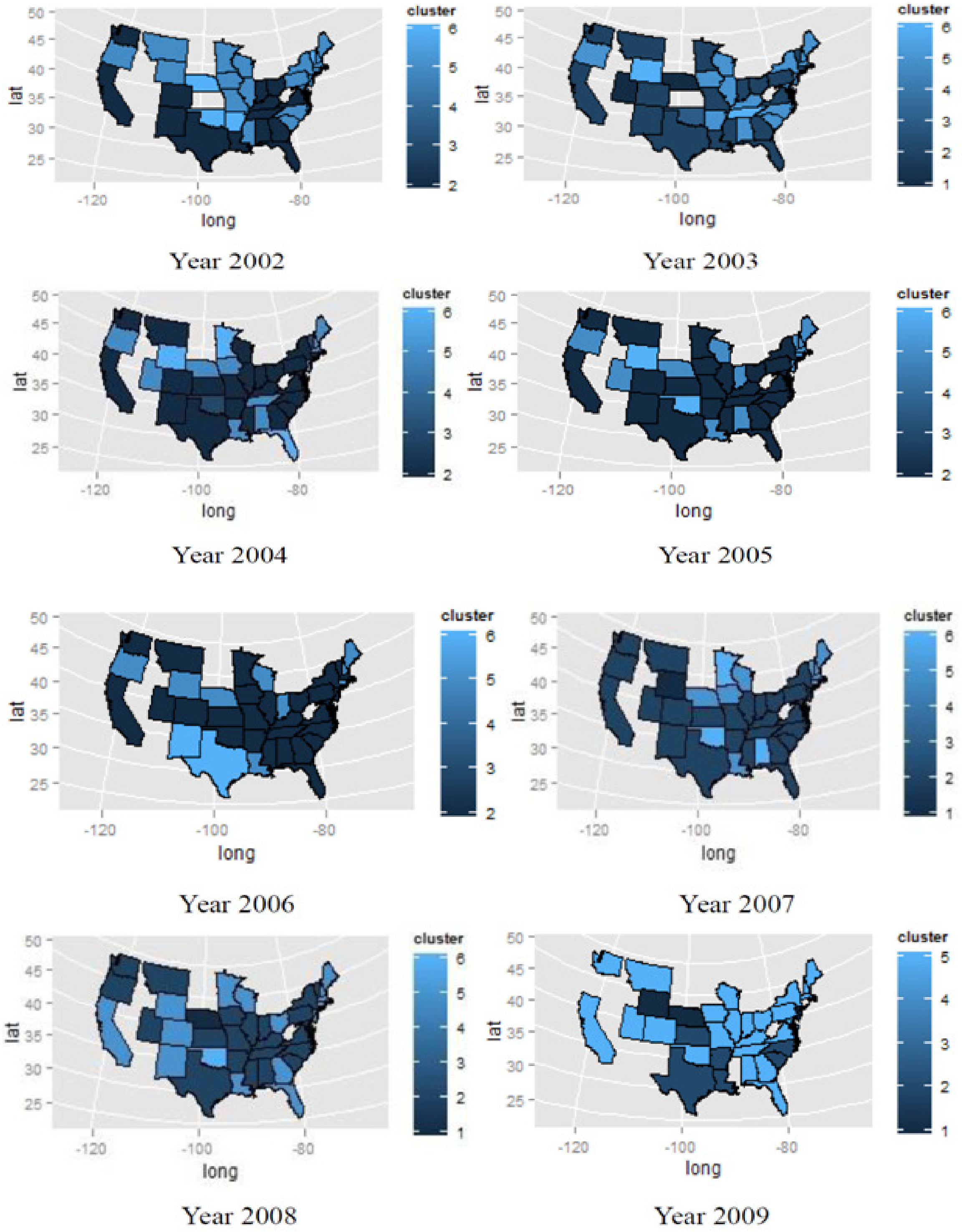


Figure 6. WithinSS for quantiles data (public schools with out-of-state tuition)

(2) Maps for private schools



Year 2002

Year 2003

Year 2004

Year 2005

Year 2006

Year 2007

Year 2008

Year 2009

Figure 7. Clustering maps for private schools using quantiles for all years

(3) Maps for public schools with out-of-state tuition

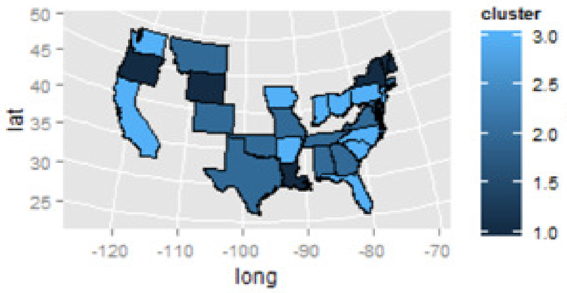
Even though there are 6 clusters in total for public schools with out-of-state tuition, there is only 1 cluster for each year. So we do not draw the figure. After checking the original data, we find out that this is because the number of public schools is too small.

4.1.2 Using Quantiles in Every Single Year

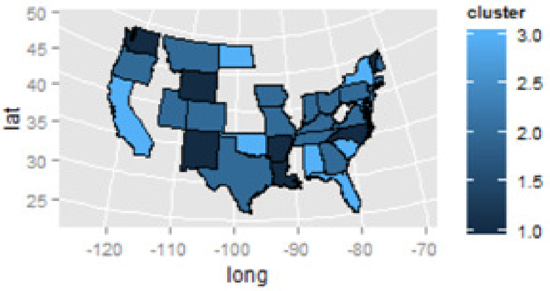
(1) Choose the number of clusters

Repeat the same procedure as the above section, we choose to use 3 clusters for both private schools and public schools without-state tuition.

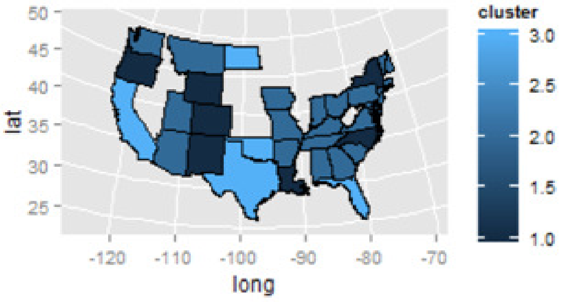
(2) Maps for public schools with out-of-state tuition



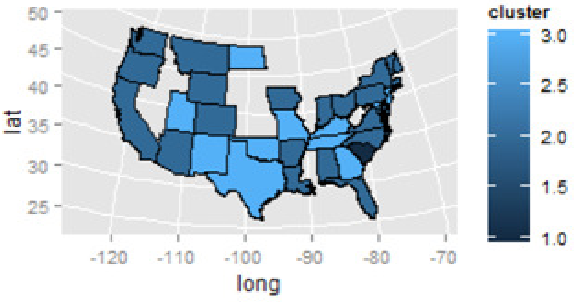
Year 2002



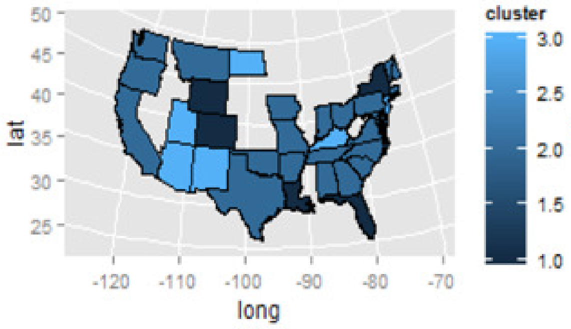
Year 2003



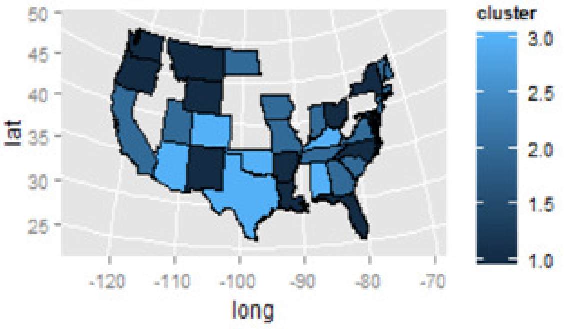
Year 2004



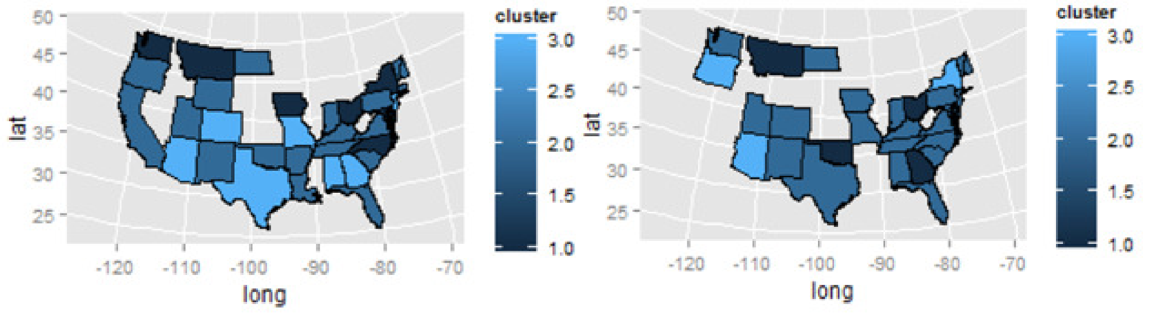
Year 2005



Year 2006



Year 2007



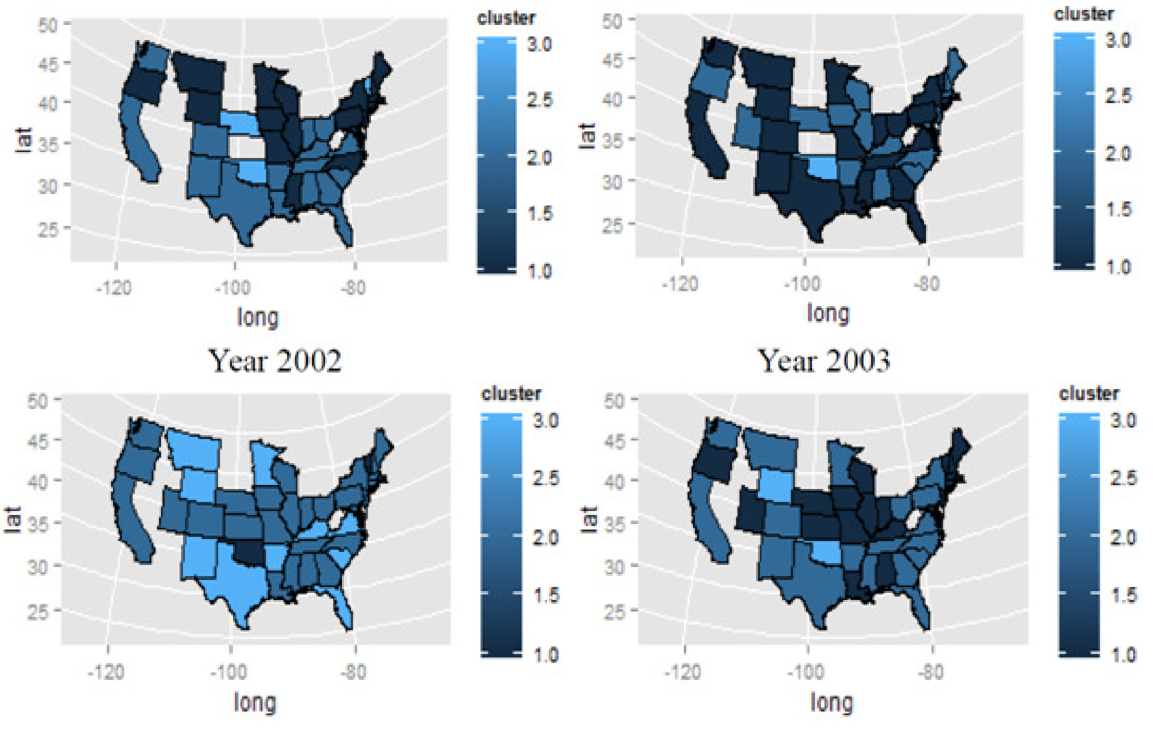
Year 2008

Year 2009

Figure 8. Clustering maps for public schools with out-of-state tuition using quantiles in every single year

From Figure 8 we can see that for the northeast coast, there is a sharp rise in year 2003 and year 2009 separately. This maybe because the Iraq War starts from March, 2003 and costs the government too much money, resulting in a budget cut for universities. Meanwhile, the financial crisis starts from 2008, causing a shortage of funds in universities. So tuition growth rates become higher. The similar situation also occurs in southern states, where the tuition growth rates are very high after year 2003. By contrast, tuition growth rates almost remain steady in west coast.

(3) Maps for private schools

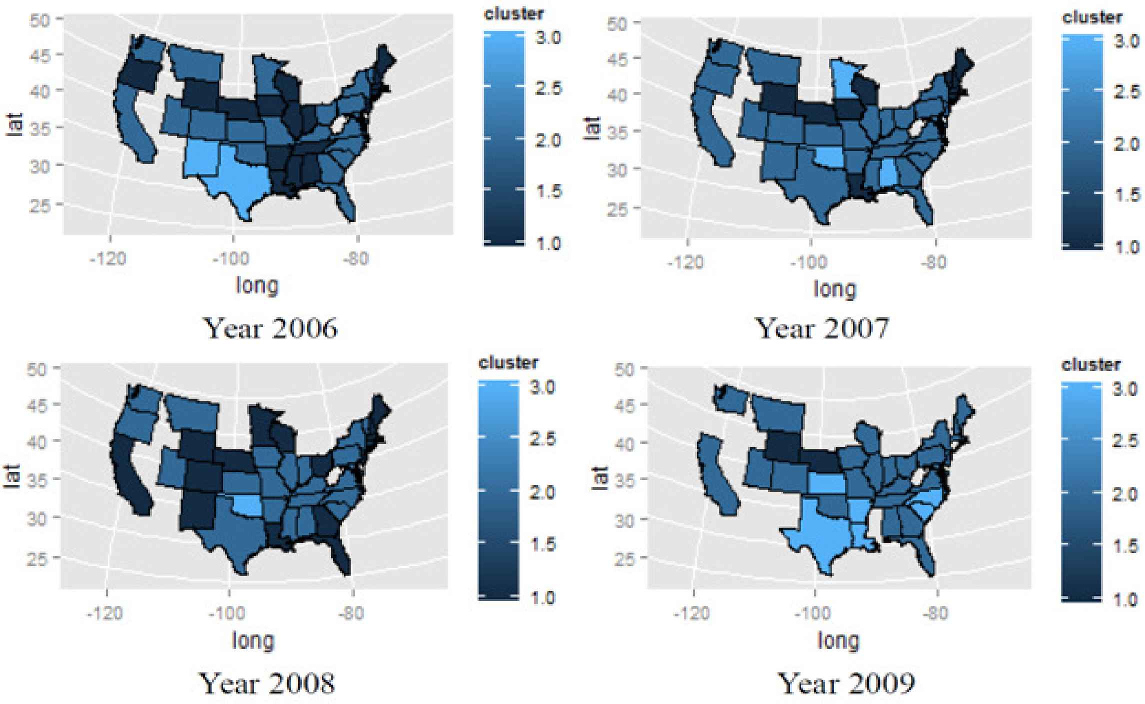


Year 2002

Year 2003

Year 2004

Year 2005



Year 2006

Year 2007

Year 2008

Year 2009

Figure 9. Clustering maps for private schools using quantiles in every single year

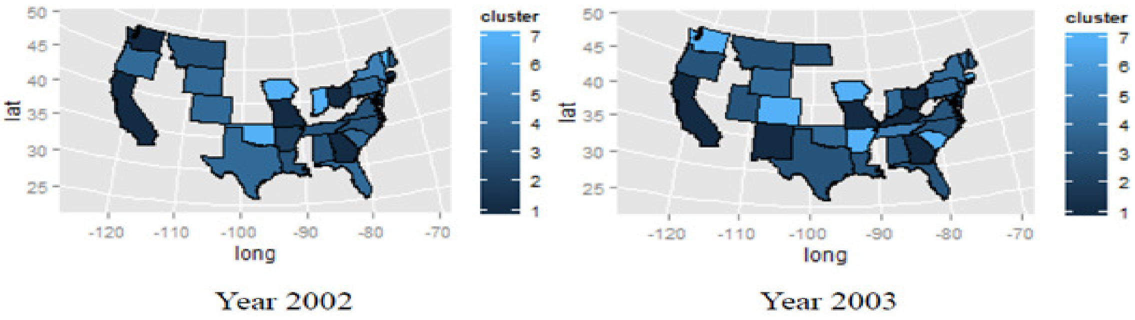
In Figure 9, it can be seen that from 2002 to 2009, the color of the middle area is first light, then it becomes darker and darker year by year, which indicates the tuition rates increase more and more slowly.

4.1.3 Using All Tuition Growth Rates

(1) Choose the number of clusters

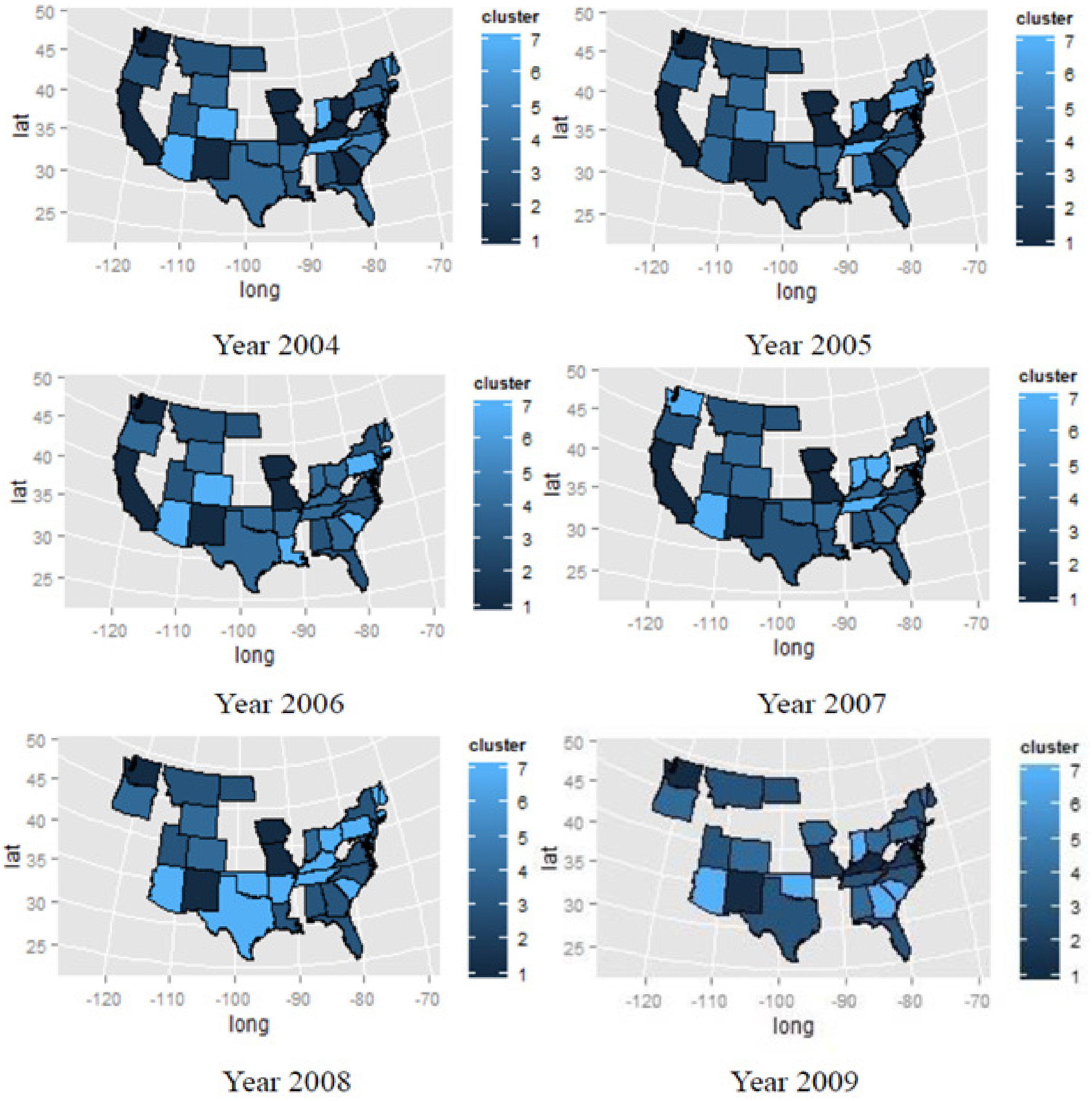
We repeat same procedures as above, and we finally choose to make 7 clusters for both public schools with in-state tuition and public schools with-out-of-state tuition. And we do not do clustering analysis for private schools. To save place, this paper just shows maps for public schools with in-state tuition.

(2) Maps for public schools with in-state tuition



Year 2002

Year 2003



Year 2004

Year 2005

Year 2006

Year 2007

Year 2008

Year 2009

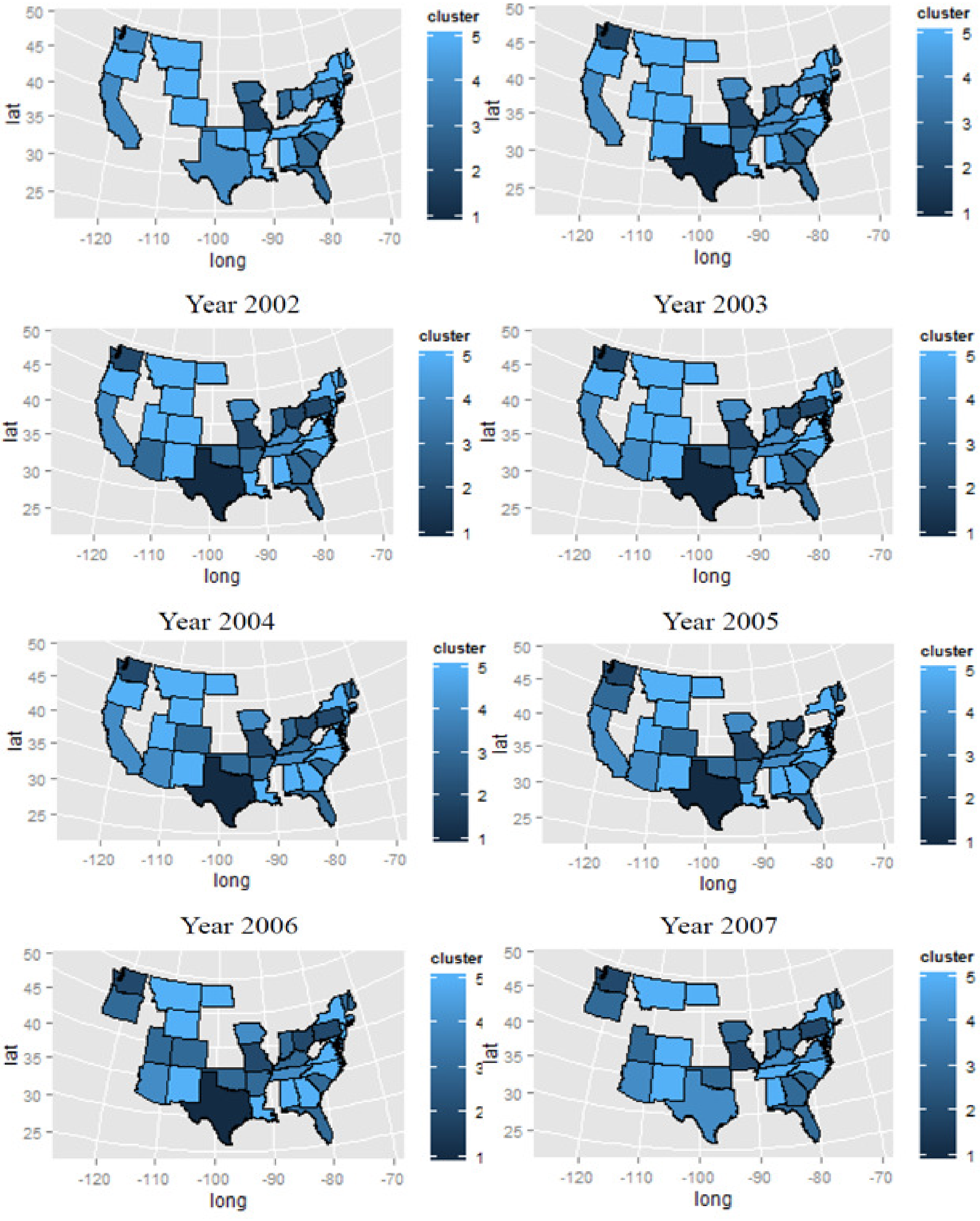
Figure 10. Clustering maps for public schools with in-state tuition using all tuition growth rates

4.1.4 Using Raw Tuition Fees

(1) Choose the number of clusters

The same procedures are repeated as above, 5 clusters are finally chosen for both public schools with in-state tuition and public schools with-out-of-state tuition using all raw tuition fees in the dataset. And we do not do clustering analysis for private schools. To save place, this paper just shows maps for public schools with in-state tuition.

(2) Maps for public schools with in-state tuition



Year 2002

Year 2003

Year 2004

Year 2005

Year 2006

Year 2007

Year 2008

Year 2009

Figure 11. Clustering maps for public schools with in-state tuition using raw tuition fee

4.2 Regression Models

To make a more comprehensive analysis, we first winsorizes outliers because there are many unreasonable outliers in various variables, shown in Figure 11. Then we create dummy variables for academic year and use state and clusters as control variables. Also, we discover a collinear phenomenon existed between "employee to student ratio" and "faculty to student ratio". In addition, our regression model also uses some other variables including total revenue, expenditures, restricted revenue, states, academic years and so on. This paper uses linear regression, regression tree, decision tree and random effect model to study the rising tuition rates in the United States.

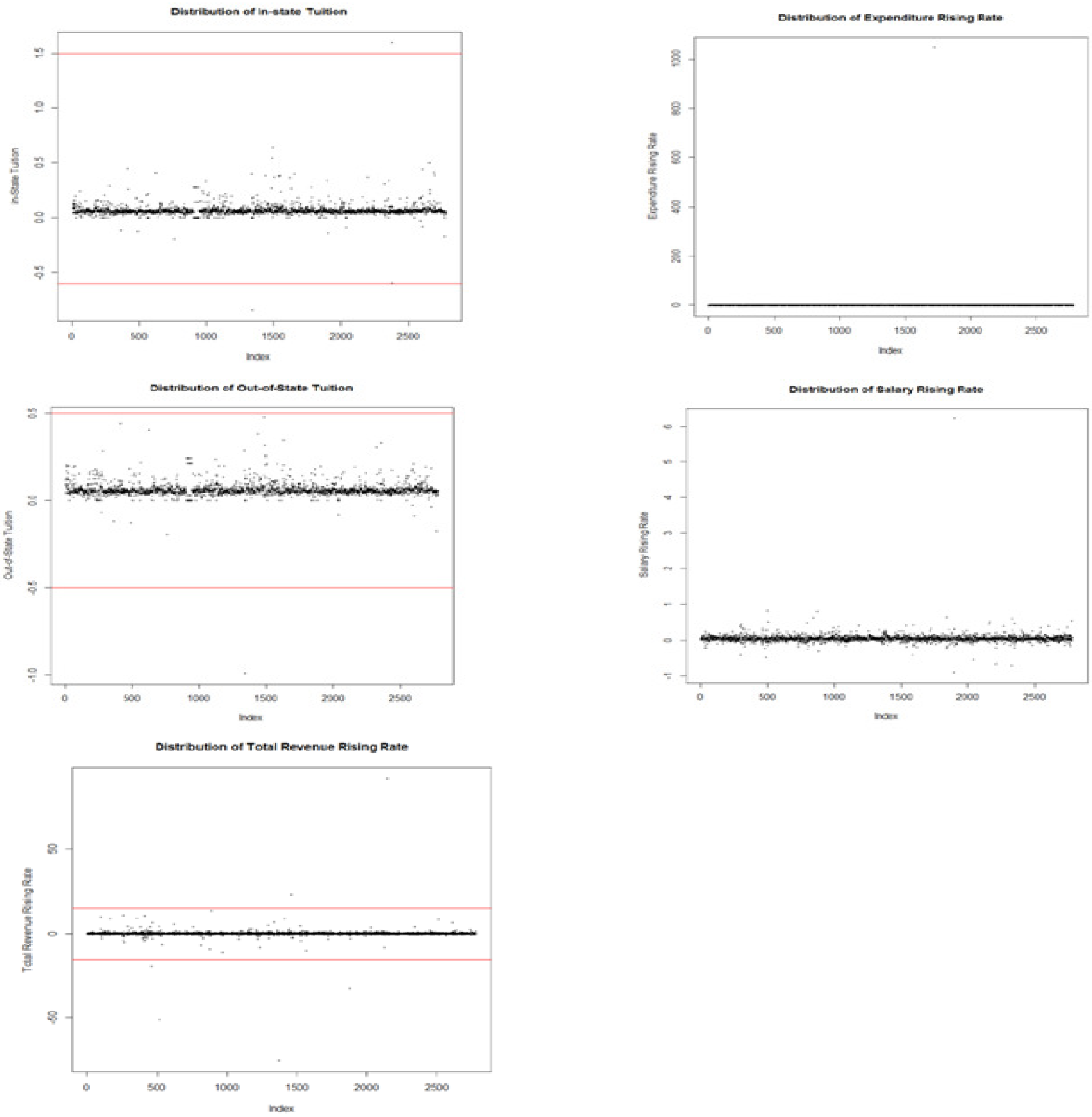
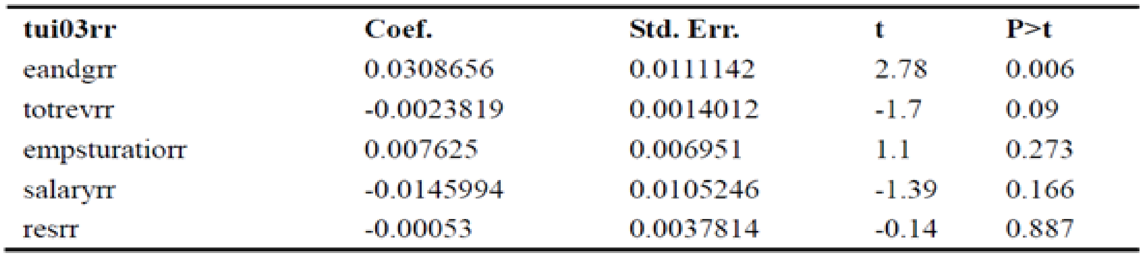


Figure 12. Data distributions and outliers for various variables

(1) Linear regression results

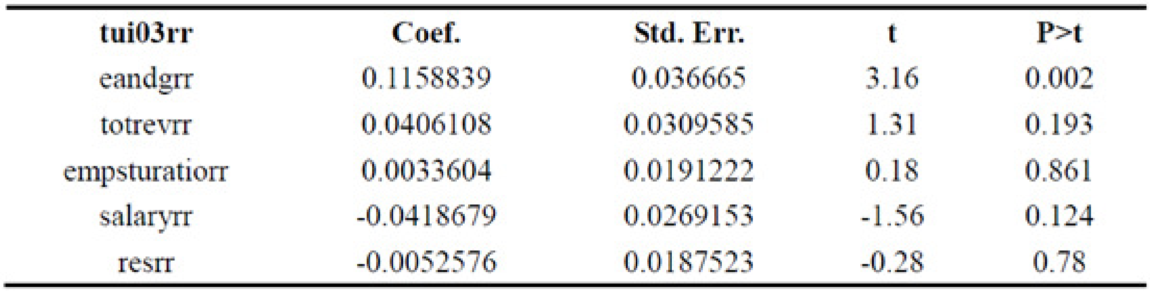
All models' R-squared is relatively small, which means this model does not fit the dataset very well. In addition, the predicted results using this linear regression model with the test dataset just capture a part of the trend about the tuition growth rates. The detailed regression results are shown in the following Tables.

Table 1. Regression results for all schools



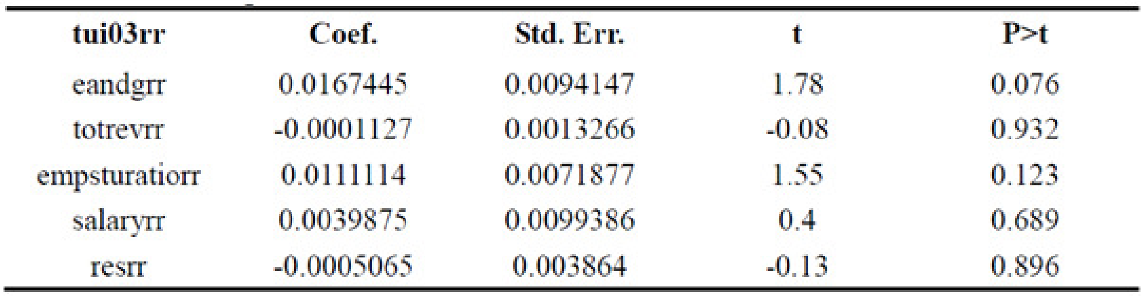
tui03rr	Coef.	Std. Err.	t	P>t
eandgrr	0.0308656	0.0111142	2.78	0.006
totrevrr	-0.0023819	0.0014012	-1.7	0.09
empsturatiorr	0.007625	0.006951	1.1	0.273
salaryrr	-0.0145994	0.0105246	-1.39	0.166
resrr	-0.00053	0.0037814	-0.14	0.887

Table 2. Regression results for private schools



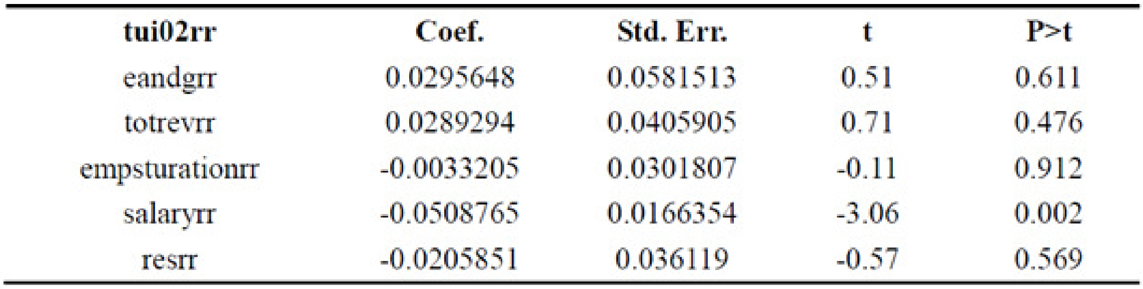
tui03rr	Coef.	Std. Err.	t	P>t
eandgrr	0.1158839	0.036665	3.16	0.002
totrevrr	0.0406108	0.0309585	1.31	0.193
empsturatiorr	0.0033604	0.0191222	0.18	0.861
salaryrr	-0.0418679	0.0269153	-1.56	0.124
resrr	-0.0052576	0.0187523	-0.28	0.78

Table 3. Regression results for public schools with out-of-state tuition



tui03rr	Coef.	Std. Err.	t	P>t
eandgrr	0.0167445	0.0094147	1.78	0.076
totrevrr	-0.0001127	0.0013266	-0.08	0.932
empsturatiorr	0.0111114	0.0071877	1.55	0.123
salaryrr	0.0039875	0.0099386	0.4	0.689
resrr	-0.0005065	0.003864	-0.13	0.896

Table 4. Regression results for public schools with in-state tuition

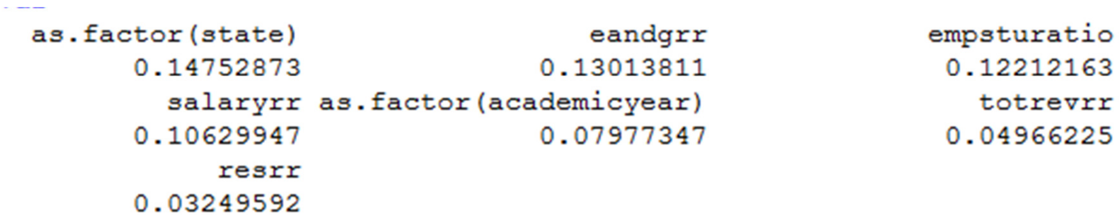


tui02rr	Coef.	Std. Err.	t	P>t
eandgrr	0.0295648	0.0581513	0.51	0.611
totrevrr	0.0289294	0.0405905	0.71	0.476
empsturationrr	-0.0033205	0.0301807	-0.11	0.912
salaryrr	-0.0508765	0.0166354	-3.06	0.002
resrr	-0.0205851	0.036119	-0.57	0.569

(2) Regression tree results

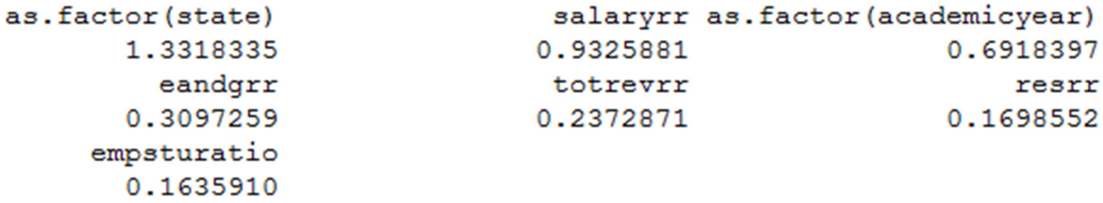
The regression trees results are shown as follows:

Variable importance for private schools:



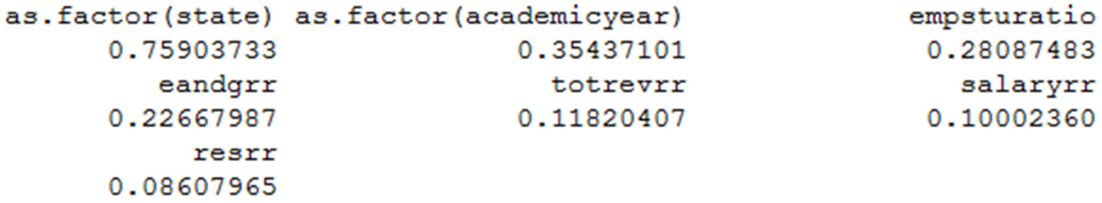
```
as.factor(state)                    eandgrr            empsturatio
      0.14752873                 0.13013811             0.12212163
        salaryrr as.factor(academicyear)                 totrevrr
      0.10629947                 0.07977347             0.04966225
           resrr
      0.03249592
```

Variable importance for public schools with in-state tuition:



```
as.factor(state)                  salaryrr as.factor(academicyear)
       1.3318335                 0.9325881               0.6918397
         eandgrr                  totrevrr                   resrr
       0.3097259                 0.2372871               0.1698552
     empsturatio
       0.1635910
```

Variable importance for public schools with out-of-state tuition:



```
as.factor(state) as.factor(academicyear)            empsturatio
      0.75903733              0.35437101             0.28087483
         eandgrr                totrevrr               salaryrr
      0.22667987              0.11820407             0.10002360
           resrr
      0.08607965
```

(3) Decision tree and random effect model

To account for the random effects in the analysis of tuition growth rates, we also propose a decision tree and random effect model in this paper. It is shown as follows:

$$y_i = f(\mathrm{X}_1, \cdots, \mathrm{X}_p) + \mathrm{Z}_i b_i + \varepsilon, \qquad i = 1, \cdots, N \tag{1}$$

Fixed effects are described using decision tree model $f(\mathrm{X}_1, \cdots, \mathrm{X}_p)$ and the random effect is considered using $\mathrm{Z}_i b_i$. Though results of this model are very similar to the results in the regression tree section, this model can combine the merits of a tree-based model and a random effect model, which means that this model is capable of analyzing dataset with irregular data, tolerating variable selection bias and reducing computational cost.

5. CONCLUSION AND DISCUSSION

After carefully performing data cleaning and model analysis, several conclusions can be drawn in this paper. It has been emphasized that the tuition growth rates have a sharp increase from 2003-2004 for public schools with in-state tuition, a sharp increase from 2004-2005 for public schools with out-of-state tuition, and a sharp increase from 2008-2009 for private schools. We deduce that these phenomenon are correlated with 2003 Iraq War, and 2008 Financial Crisis. From the clustering analysis, we find out that for private schools the tuition growth rates increase year by year for nearly all states. And for public schools, tuition growth rates of public schools on the west coast grow more slowly than that of public schools on the east coast. From the regression models, it can be demonstrated that for all types of schools, both expenditures and total revenues affect the tuition growth rate significantly. More specifically, the tuition growth rate is positively related with the expenditure while negatively related with the total revenue.

REFERENCES

[1] Ehrenberg, Ronald G, (2012) "American higher education in transition", The Journal of Economic Perspectives, pp193-216.

[2] Bell, Daniel, (1976) "The coming of the post-industrial society", The Educational Forum, Vol. 40, No. 4. Taylor & Francis Group.

[3] Garrett, Thomas A, and William Poole, (2006) "Stop paying more for less: ways to boost productivity in higher education", The Regional Economist, Vol. 4, No. 9.

[4] Chiodo, Abbigail, and Michael T. Owyang, (2003) "Financial aid and college choice", Economic Synopses, 2003-08-03.

[5] Heller, Donald E, (1996) "Tuition Prices, Financial Aid, and Access to Public Higher Education: A State-Level Analysis".

[6] Betts, Julian R., and Laurel L. McFarland, (1995) "Safe port in a storm: The impact of labor market conditions on community college enrollments", Journal of Human Resources, pp741-765.

[7] Jackson, Gregory A., and George B. Weathersby, (1975) "Individual demand for higher education: A review and analysis of recent empirical studies", The Journal of Higher Education, pp623-652.

[8] Heckman, James J., Lance Lochner, and Christopher Taber, (1998) "General equilibrium treatment effects: A study of tuition policy", No. w6426, National Bureau of Economic Research.

[9] Campbell, Robert, and Barry N. Siegel, (1967) "The demand for higher education in the United States, 1919-1964", The American Economic Review, pp482-494.

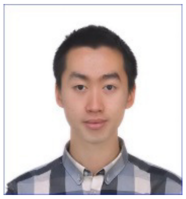
[10] Sewell, William H, (1971) "Inequality of opportunity for higher education", American Sociological Review, Vol. 36, No. 5, pp793-809.

[11] Heckman, James J, (1997) "Instrumental variables: A study of implicit behavioral assumptions in one widely used estimator", Journal of Human Resources, Vol. 32, No. 3.

[12] U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Data Source Website: https://nces.ed.gov/ipeds/deltacostproject.

AUTHOR

Long Cheng is currently working as the COO and research scientist at Kiwii Power Technology Corporation, Troy, NY, USA. Before that, he worked as a data scientist at Rang Technologies Inc, Piscataway, NJ, USA. He received his Master’s Degrees in both Electrical Engineering and Applied Mathematics from Rensselaer Polytechnic Institute, Troy, NY, USA in May 2015 and his B.S. in Electrical Engineering and Automation from Tianjin University, Tianjin, China in July 2013. His research interests include machine learning, data mining, signal processing and smart grids.



Chenyu You is currently pursuing his B. S. in Electrical Engineering with a minor in Mathematics from Rensselaer Polytechnic Institute, Troy, NY, USA. His research interests are machine learning, data mining, statistical signal processing and mathematical modelling.