The Big Data Based Effect of Household Wealth on Elementary Student Achievement (Take Shanghai for Example)

Xuesheng Qian, Yifeng Xu, Jing Zhang, and Wei Zhao

***Abstract*—The students' academic achievement is the globally general criteria to evaluate effectiveness of the education, while in addition to the students' own talents, efforts and educational resources, the family's Socioeconomic Status (SES) plays a crucial role. The previous studies target the indicators of SES scale, while this thesis initially breaks the drawbacks and limitations of the conventional scale methodology in the research of household wealth survey, utilizing the Big Data Theory to carry out a household wealth survey from 27k students' families of 56 primary schools in a district of Shanghai. Eventually, the thesis implements the analysis of the household wealth's effect on the academic achievement.**

***Index Terms*—Big data, family's welfare, SES, student's score, web crawler.**

I. Introduction

A. SES Influence on Academic Achievement

The students' academic achievement has always been one of the essential criteria to assess the education's effectiveness of one country. Apart from the students' talent, hardworking and education resources, the family's SES plays a crucial impact on the students' academic development. Since 1966 when Coleman published the report on equal opportunities in education research, many scholars carried out bountiful researches on the children's academic achievement under the impact of SES. Professor Hu Yongmei, in the study of western China primary schools, found that family's SES had a remarkably influence on the academic score. The Chinese and mathematics scores of the students from high SES families significantly excelled than the others [1]. With the development of China's economy and the each common family's wealth, social mobility has become a leading issue recently. Program for International Student Assessment (PISA) pointed out that social stratification had a convincing impact on the educational outcome. The study of Fang Changchun took the samples covering both the rural and urban schools, general and vocational schools and he found that the entrance examination score for high school was expressively associated with the family's social class. Not just limited in the academic development, some studies stated that the SES could also affect students' health, cognition throughout period from birth to adult [2]. Xue Haiping and Wang Rong applied the Multilevel Model to analyze and came to the conclusion that the family's socioeconomic background had a compellingly positive correlation with the students' primary and middle school mathematics score [3]. Xue's study divided the income levels according to whether the family owned a computer. And the applicability of this criteria is still debatable.

Via researching volumes of literature, Selcuk R. Sirin pointed out that the divergent indicators of SES might lead the different influence on academic achievement, which might cause inconsistent conclusion [4]. When the study came to the diverse school or personal levels, the criteria of measuring the SES was not equal. Other existing studies found that the relationship between the SES and academic achievement on the school level varied from countries to countries [5]. Therefore, the key indicator of the SES is still controversial and undetermined, despite of the current findings of SES and their relevant academic accomplishments.

The thought that family's SES has a significant impact on children's academic achievement has become the consensus of the academic community, although the depth, the way of the impact and the influence discrepancy of different disciplines still needs further studies to complete. Ding Yanqing and Xue Haiping took a analysis on the factors influencing college entrance examination scores of the public high school of Kunming in 2006 and they found out that the family's SES had no convincing impact on the scores and some results depicted that the average monthly household income had a negative effect. The possible explanation shows that in the high school entrance examination, the students with high family' SES are more likely to get the entrance opportunity compared to those with the same academic level. Therefore, the students from the less rich families should be super outstanding to enter the high school education. The impact of family's SES on the high school students' score is not easily determined and since the other relevant factors are not smoothly controlled, the conclusion appears much unreliable.

In addition to the direct researches on the correlation between SES and student' academic achievement, a large number of Mediating Effects Model researches emerged in recent years (Cheng Yinghua & Mao Yaqing, 2016; Shi Leishan, Chen Yingmin & Hou Xiu, 2013). The study found that as SES's Mediating variables, learning inputs (focus, dedication and vigor) and social emotional competence (self-awareness, self-management, others' cognition, others'

Manuscript received August 12, 2016; revised January 22, 2017.

Xuesheng Qian is with the Antai College of Economics and Management, Shanghai Jiaotong University, China (e-mail: e_qianxuesheng@126.com).

Yifeng Xu, Jing Zhang, and Wei Zhao are with Psylife Consulting Co.,Ltd, Shanghai, China.

doi: 10.18178/ijiet.2017.7.11.986

management, group awareness and group management) had significant impacts on student's academic achievement.

Through the research and analysis of 27670 fourth grade primary school students from 56 ordinary schools in one district of Shanghai, this thesis aims to analyze the impact of family's SES on the different disciplines in the relatively developed areas, targeting to supply previous studies and provide scholarly references to successors.

B. SES Measurement Methodology

Although SES is seen as one core indicator in many studies, there still exists controversy about its inner concept, empirical measurement and standard in various countries. The measurement methodology of SES can be broken down into Self-Reporting method (or Filling method), Interval Selection method. But the inconsistent classification, interval range, size settlement and the defects of measurement and calculation, etc. may result in in-consistence, contradictory and unclear interpretation. When the same samples are applied with diverse methods to assess, the outcome may appear significant discrepancy [6]. In conventional research, the indicators of SES Index relies on the amount of social resources which the family can acquire and control, such as parents' income, educational level, occupational prestige or some other focusing on the family's possession to indirectly obtain the SES level, for instance, televisions, refrigerators, washing machines and computers, etc. [7]. The applicability of this measurement methodology varies in the unbalanced economic development level of different regions, easily leading large system errors and discrepancies of the conclusion.

In various SES evaluation indicators, family's economic indicators are always the key information. Simultaneously, the family's economic condition, as the core indicator of SES, also affects other related indicators. Many empirical studies show that the family's income has certain correlation with other factors (Bollen, Glanville & Stecklov, 2001; Hauser & Huang, 1997). The family's income surveys in the past usually require students or parents to fill in the income value directly. Although the income is the most representative indicator to demonstrate the family's economic status, due to the high degree of privacy, the real data fails to be collected. Besides, the survey's content mainly be filled out by students and they incline to give the distorting data or even miss filling some information since they do not know the real income of their parents. Consequently, the no-response rate with regard to the family's income question is nearly 15% through the questionnaire form method to collect the data [8].

Under the special background of Chinese Market Economy System Reform, "Gray Income" comes into appearance. Li Zhining in his *A Chinese Economist' s Worry*, pointed out that there were 1.5 times payment's vacuum in Chinese urban population income in 2002, implying there existing 1.5 trillion "gray income" [9]. Wang Xiaolu calculated in 2009 that the 14 trillion disposable income of urban residents was not reflected in the income statistics, among which the 80% hidden income was concentrated in top 20% high-income urban households and 60% in the top 10% [10]. Currently, the proportion has already far exceeded the original, therefore in China especially in the districts with high level of economic development like Shanghai, the income cannot fully reflect the social and economic status of the residents.

In contrast, China's central bank cut interest rate and required reserve ratio in recent years and the bank deposit interest rate was far from keeping up the price index. Other investment channels became narrow and full of risk, leading money flowing into the the real estate market and pushing up the prices in Chinese cities [11]. Especially in Shanghai, the real estate price has high correlation with the family's economic level, therefore the prices can be treated as the new distinctive indicator.

In data collection, this thesis applies the more objective Big Data technology to obtain the required information, eliminating the defects like data loss, data error, high cost and low efficiency in the artificial methodologies. Through the introduction of the new techniques, the Big Data method can supply the existing study and provide a further reference for further researches.

II. Data Collecting

A. The Introduction of Big Data Methodology

With the improving maturity and development of the market, the big data theory, relying on its market demand and commercial value, is becoming the novel engine to speed up the information industry and even the whole economic and social change [12]. Statistically, the China's big data industry market size is over 110.56 billion, with a 44.15% increase compared to 2014, where the big data infrastructure, big data software and the application share a proportion of 64.53%, 25.47% and 10% respectively [13]. In the era of big data, the main trends can fall into three aspects: data becoming an asset, vertical integration, Internet-Based development. And these accelerate the Enterprise Innovation and promote the information industry and the whole economic and social change on a global scale [14]. IDC's report *2020 Digital Universe: The Big Data and the Larger Digital Shadow Will Achieve the Fastest Growth in the Far East Places* predicts that from 2013 to 2020, the digital universe will be doubled every two years.

In the field of education, the big data theory can bring novel ideas and perspectives, making it possible to follow and track the individuals' development and demand. Optimizing the allocation of educational resources, improving the teaching quality, promoting the educational equity, analyzing and forecasting teaching behaviors and students' learning behaviors, developing personality of students will be fully specified with appropriate, accurate and reliable theory. Under the support of the big data, the education policies will appear forward-looking and be properly navigated.

In the era of big data, the network information grows explosively in the form of both structured and unstructured data. Therefore the demand of big data promotes the popularization and development of web crawler. Currently, due to the abilities of searching and gathering specific information, the flexibility of handling multimedia information, saving hardware storage and network resources,

there exist two popular algorithms as commonly used: Universal Web Crawler and Topic Web Crawler [15].

The basic flow of the web crawler is illustrated in the Fig. 1. The main object processed by the web crawler is URL, based on the Http protocol. After accessing and analyzing the web pages, this thesis applies the regular expressions to match the HTML tags for text extraction and then stores them into the database.

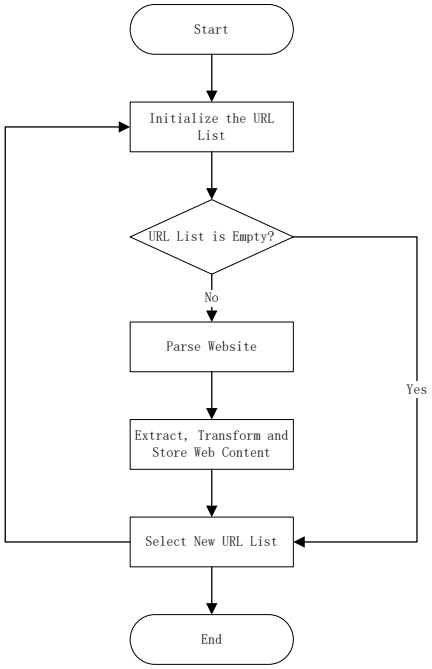


Fig. 1. Web crawler's basic flowchart.

B. Price Information Crawling

Depending on the the web crawler technology, this thesis matches the each real estate price of the student's family with the residence address from the Anjuke Website. Anjuke, established in 2007, is a leading real estate information service platform, having set up 50 branches in China with its service in 500 cities. Based on the statistics of the first season of 2016, the number of web browsing per month is more than 100 million. The Anjuke App has been installed in the mobile terminals about 120 million times, accounting for 70% market volume, ensuring the reliability and validity of the crawled information.

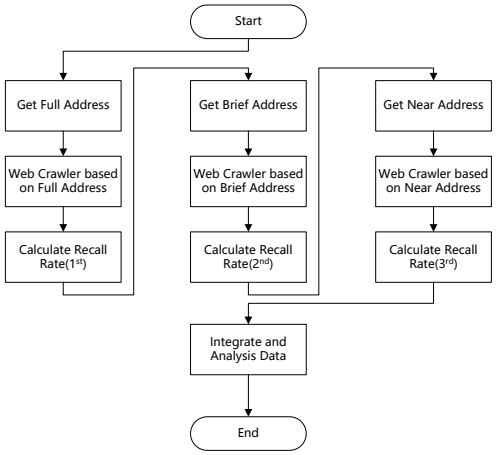


Fig. 2. Price gathered by web crawler algorithm flowchart.

The web crawler's specific algorithm flow is shown in Fig. 2:

Web crawler flowchart description:

1) Selecting students' address, excluding incomplete and improper address;
2) Crawling the price information from the Anjuke Website according to the selected addresses;
3) Calculating the recall rate;
4) According to the results of the first crawling process, streamlining the address information;
5) Crawling price information secondly;
6) Calculating the recall rate secondly;
7) In accordance with the results of the second, optimizing the strategy;
8) Crawling price information thirdly;
9) Calculating the recall rate thirdly.

The final crawler result is demonstrated in Table I:

TABLE I: SUMMARY OF WEB CRAWLER INFORMATION

Total People	The No. Found	Recall Rate	Means	Standard Dev.
31662	27670	87%	37554.75	12033.05

C. Student's Academic Achievement

Selecting 27670 nine-year and ordinary primary school students from 56 schools in a representative evenly-distributed-population district of Shanghai, the study analyzes Chinese, Mathematics and English subjects. Since the three courses represent three different capabilities and the previous studies state that the family's SES has diverse impacts on the three courses, this thesis not only evaluates the comprehensive score, but also assesses the each subject under the SES's influence.

The elementary school can be broken down into nine-year and ordinary primary school. And the real estate price falls into three level: 20K~30K, 30K~40K, 40K~50K. The proportion of household registration and gender ratio can be divided into 4 or 3 levels. As listed in Table II. Household Registration Proportion

TABLE II: SCHOOL DISTRIBUTION IN DIFFERENT LEVELS ACCORDING TO THE INDEPENDENT VARIABLES

Argument	Hierarchy				Total
School Property	Nine-year	Ordinary			
	13	43			56
Average Price	20k~30k	30k~40k	40k~50k		
	13	24	19		56
Household Registration Proportion	0.12~0.32	0.32~0.52	0.52~0.72	0.72~0.92	
	9	14	22	11	56
Gender Ratio	0.45~0.50	0.50~0.55	0.55~0.60		
	12	27	17		56

III. ANALYSIS AND RESULT

A. Overall Score Analysis

Through the Full Order Multivariate Analysis of Variance, the Analysis of Variance table is illustrated below:

As illustrated in the multivariate analysis of variance table, the most significant factor of the overall score is the proportion of household registration, where p value is 0.017, then followed by the interaction of average real estate price

and the proportion of sex ratio, with 0.047. Other variables do not reach convincing level, statistically.

TABLE III: MULTIVARIATE ANALYSIS OF VARIANCE OF OVERALL SCORE

	Df	Pillai	approx F	num Df	denDf	Pr(>F)
avg_hp	2	0.279	1.621	6	60	0.157
cr_prop	3	0.565	2.396	9	93	0.017*
gender_prop	2	0.146	0.786	6	60	0.584
avg_hp: cr_prop	6	0.435	0.876	18	93	0.608
avg_hp: gender_prop	3	0.488	2.009	9	93	0.047*
cr_prop: gender_prop	4	0.233	0.654	12	93	0.790
avg_hp: cr_prop: gender_prop	3	0.345	1.341	9	93	0.227
Residuals	31					

*Sig.: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1*

P-value is calculated by the robustest Pillai's criterion statistics

To further investigate the relationship between the each argument and dependent variable, the following paragraphs give the multivariate analysis of variance to each subject.

B. Chinese Score Analysis

Through the Full Order Univariate Multivariate Analysis of Variance of Chinese, the Analysis of Variance table is illustrated as follows:

TABLE IV: VARIANCE ANALYSIS OF CHINESE SCORE

	Df	Sum Sq	Mean Sq	F_value	Pr(>F)
avg_hp	2	19.57	9.787	2.355	0.112
cr_prop	3	12.03	4.009	0.965	0.422
gender_prop	2	2.92	1.462	0.352	0.706
avg_hp:cr_prop	6	20.19	3.365	0.810	0.570
avg_hp:gender_prop	3	58.83	19.610	4.720	0.008**
cr_prop:gender_prop	4	2.23	0.559	0.134	0.968
avg_hp:cr_prop: gender_prop	3	15.55	5.184	1.248	0.309
Residuals	31	128.80	4.155		

*Sig.: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1*

Based on the Analysis of Variance table, in regards of Chinese scores, the interaction between the average real estate price and sex ratio is significant at the 99% confidence level (F(df=3)=4.72, p=0.008). Other variables do not reach statistically significant levels.

C. Mathematics Score Analysis

Similarly, through the Full Order Univariate Multivariate Analysis of Variance of Mathematics, the Analysis of Variance table is illustrated as follows:

TABLE V: VARIANCE ANALYSIS OF MATHEMATICS SCORE

	Df	Sum Sq	Mean Sq	F_value	Pr(>F)
avg_hp	2	21.81	10.90	1.604	0.217
cr_prop	3	50.77	16.92	2.490	0.078.
gender_prop	2	18.12	9.06	1.333	0.278
avg_hp:cr_prop	6	38.22	6.37	0.937	0.483
avg_hp:gender_prop	3	105.00	35.00	5.150	0.005**
cr_prop:gender_prop	4	4.34	1.09	0.160	0.957
avg_hp:cr_prop:gender_prop	3	40.03	13.34	1.964	0.140
Residuals	31	210.67	6.80		

*Sig.: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1*

As illustrated from the above, in regards of Mathematics score, the proportion of household registration is at 90% level(F(df=3)=2.49, p=0.078). The interaction of average price and sex ratio is at 99.9% confidence level(F(df=3)=5.15, p=0.005). Other variables do not reach statistically significant levels.

D. English Score Analysis

Similarly, through the Full Order Univariate Multivariate Analysis of Variance of English, the Analysis of Variance Table VI is illustrated as follows:

TABLE VI: VARIANCE ANALYSIS OF ENGLISH SCORE

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
avg_hp	2	73.49	36.75	5.109	0.0120*
cr_prop	3	200.48	66.83	9.292	0.0002***
gender_prop	2	28.67	14.34	1.993	0.1533
avg_hp:cr_prop	6	47.91	7.99	1.110	0.3788
avg_hp:gender_prop	3	122.11	40.70	5.660	0.0033**
cr_prop:gender_prop	4	38.77	9.69	1.348	0.2746
avg_hp:cr_prop: gender_prop	3	24.45	8.15	1.133	0.3509
Residuals	31	222.95	7.19		

*Sig.: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1*

As illustrated in the Table 6, in regards of English score, the average price is significant at 95% confidence level (F(df=2)=5.109, p=0.012). The proportion of household registration is at 99.9% level(F(df=3)=9.292, p=0.0002). The interaction of average price and male to female ratio is at 99.9% confidence level(F(df=3)=5.66, p=0.003). Other variables do not reach statistically significant levels.

IV. RESULTS

A. The Strong Influence on English Learning

From the above statistics, the price shows the significant effect on the English score, while the Chinese and mathematics do not appear the statistically significant level, which demonstrates that family's SES can convincingly influence the student's English achievement in this district. Consequently, it is unavoidable to compare the English scores with various prices. The results are as follows:

TABLE VII: PAIRWISE COMPARISON BETWEEN THE PRICE AND ENGLISH SCORE

	diff	lwr	upr	p adj
30k~40k-20k~30k	1.747	-0.891	4.385	0.255
40k~50k-20k~30k	3.084	0.347	5.820	0.024
40k~50k-30k~40k	1.336	-1.021	3.693	0.364

It can be concluded from the above chart that the average price in the interval of 40K~50K holds the highest English score, followed by 30K~40K, in contrast to the minimum in the interval of 20K~30K. Among them, the score's discrepancy between the 40K~50K and 20K~30K is significant (95% significance level).

In the light of this, students with high family's SES have obvious advantages in terms of English achievement, in accordance with previous studies. This phenomenon can be explained for families with high economic condition are much likely to create more opportunities and provide more

resources to learn English, such as participating cram school, traveling abroad, etc. Accordingly, it can be considered that further educational resources can be drawn into the relatively backward areas in this district.

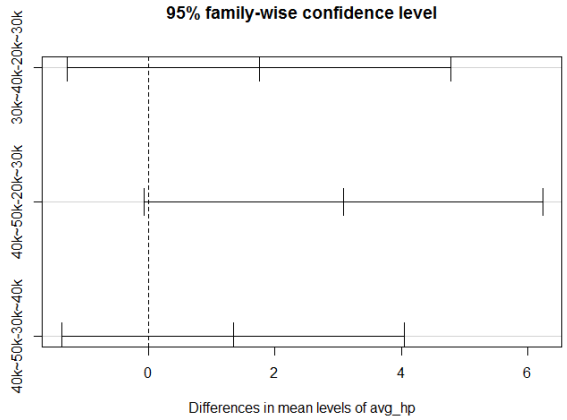


Fig. 3. 95% family-wise confidence level.

In contrast, there appears no significant effect between the high SES and students' Chinese and mathematics performance, which is contradict to some studies. But it is still possible that no convincing gap is formed in Chinese and mathematics learning for the overall deployment and implementation of basic educational policy in Shanghai, .

B. Interaction between Price And Gender Ratio

Notwithstanding no obvious discrepancy appears in the relationship between the prices and the Chinese and mathematics, via the full order analysis of variance of each subject, it is found that the interaction between the real estate price and the gender ratio is eloquent. And the interaction map is identified as follows:

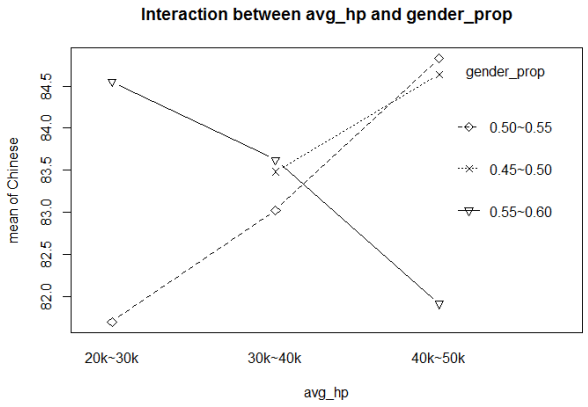


Fig. 4. Interaction between avg_hp and gender_prop.
Note: gender_prop value is equal to the number of males divided by females

The Fig.4 highlights the interaction on the Chinese subject. When the gender ratio is in the interval of 0.45~0.55, the Chinese score increases with the growth of average price. However, for the interval between 0.55 and 0.60, there exists a negative correlation.

This may reflect that the boys and girls in different economic levels behave opposed advantages. The high economic level enables the girls making a tremendous progress in Chinese, while the boys from the high economic level may be less likely to learn well in Chinese. Therefore, the high gender_prop value means the less excellent performance in Chinese.

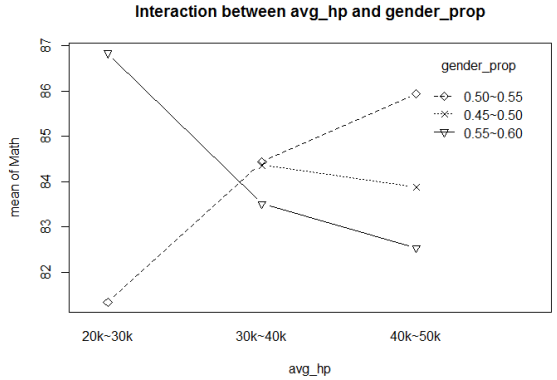


Fig. 5. Interaction between avg_hp and gender_prop.

The above figure highlights the interaction between the average price and gender ratio on the mathematics subject. When the gender ratio is in the interval of 0.50~0.55, the mathematics score increases with the growth of average price. However, for the interval between 0.45 and 0.50, the mathematics score fluctuates little with the change of average price.

This result reflects the mathematics score may increase in accordance with family's economic level in a moderate gender proportion school. In contrast, if the school shares a high gender_prop value, boys may be less good at mathematics even though they are born in high economic level families, which attracts the further reflection.

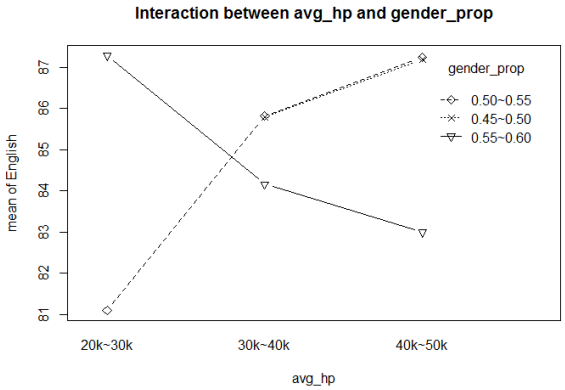


Fig. 6. Interaction between avg_hp and gender_prop.

The above figure illuminates the interaction on the English subject. When the gender ratio is in the interval of 0.45~0.50 and 0.50~0.55, the English score increases with the growth of average price. However, for the interval between 0.55 and 0.60, there exists a negative correlation.

This result is consistent with the result of the above analysis of Chinese, reflecting the same problem with it. In short, for language disciplines, the scores may appear positive correlation with the average price in the relatively low gender_prop school, while the scores are negatively correlated with the price in the schools with less girls but more boys. The reason is still under debate.

C. The Discrepancy Caused by Household Registration

In addition to the above conclusions, the schools with various proportions of household registration show convincing discrepancy in the mathematics and English performance. This may imply that the household registration is one essential factor, and therefore, the further pairwise comparisons of the variables are listed as follows:

TABLE VIII: THE COMPARISON BETWEEN MATHEMATICS PERFORMANCE AND DIVERGENT PROPORTIONS OF HOUSEHOLD REGISTRATION

	diff	lwr	upr	p adj
0.32~0.52-0.12~0.32	2.333	-0.930	5.597	0.241
0.52~0.72-0.12~0.32	3.063	0.019	6.106	0.048
0.72~0.92-0.12~0.32	3.134	-0.300	6.567	0.085
0.52~0.72-0.32~0.52	0.729	-1.907	3.365	0.882
0.72~0.92-0.32~0.52	0.800	-2.278	3.878	0.900
0.72~0.92-0.52~0.72	0.071	-2.772	2.914	0.999

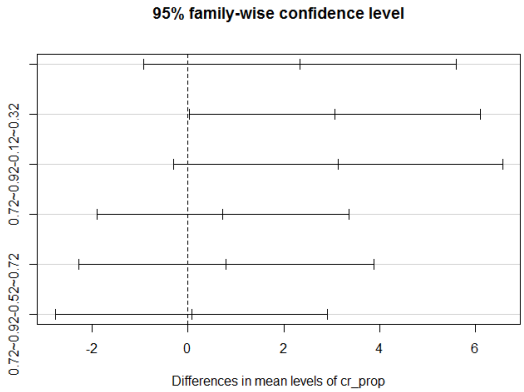


Fig. 7. 95% family-wise confidence level.

As illustrated from the above chart, the better mathematics score is co-related with the higher proportion of household registration, while the relatively lower proportion of household registration schools share less excellent scores. Among them, the average mathematics scores have convincing difference in the household registration proportion of 52% ~72% and 12%~32%(95% significance level). Also, the school household registration proportion between 72%~92% and 12 % to 32% exists compelling difference (90% significance level).

This result may reflect local students are superior to the non-local students, so it can be considered to strengthen the non-local students' mathematical and logical ability training in specific teaching process.

TABLE IX: THE COMPARISON BETWEEN ENGLISH PERFORMANCE AND DIVERGENT PROPORTIONS OF HOUSEHOLD REGISTRATION

	diff	lwr	upr	p adj
0.32~0.52-0.12~0.32	2.064	-1.510	5.639	0.424
0.52~0.72-0.12~0.32	4.135	0.802	7.468	0.009
0.72~0.92-0.12~0.32	5.429	1.669	9.189	0.002
0.52~0.72-0.32~0.52	2.070	-0.816	4.957	0.238
0.72~0.92-0.32~0.52	3.365	-0.006	6.735	0.051
0.72~0.92-0.52~0.72	1.294	-1.819	4.408	0.687

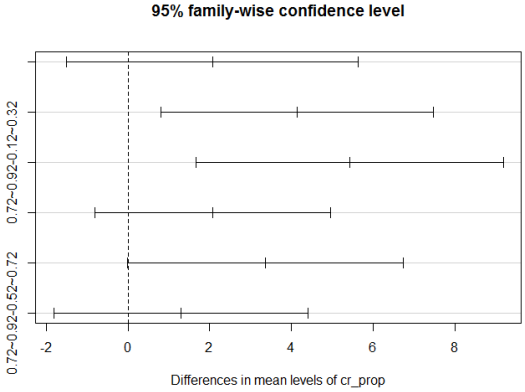


Fig. 8. 95% family-wise confidence level.

As illustrated from the above chart, the better English score is co-related with the higher proportion of household registration, while the relatively lower proportion of household registration schools share less excellent scores. Among them, the average English scores have convincing difference in the household registration proportion of 52% ~72% and 12%~32%(99% significance level). Also, the school household registration proportion between 72%~92% and 12 % to 32% exists compelling difference (99% significance level).

This result is consistent with the above analysis which can be reasoned that the local students are much likely to access more English learning resources.

V. CONCLUSION

This study initially proposes a novel big data methodology and realizes the web crawler as a further expand to the conventional questionnaires and the data reports. Through the web crawler technique to collect the lossless sample data, the conclusions can ensure the data's objectivity and authenticity, which in some points, enhances the scientificity and accuracy of the discovers. The advanced methodology for this study is a big breakthrough not only for itself, but also for the whole social science, and it is building a new bridge between traditional methodology and big data.

Via the above analysis, the following conclusions are drawn from the study: In general, household wealth plays an essential role on student achievement, especially in English study. Further more, the effect of proportion of household registration shows convincing discrepancy in terms of mathematics and English learning.

REFERENCES

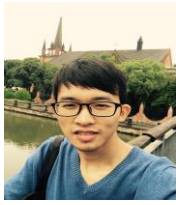
[1] Y. Hu and Y. Du, "Empirical research on the educational production function of rural primary schools in western China," *Educational Research*, vol. 07, pp. 58-67, 2009.
[2] C. Fang and X. Feng, "How distinction of social stratum affects the attainment of education: An analysis on split-flows of education," *Tsinghua Journal of Education*, vol. 05, 2005.
[3] H. Xue and R.Wang, "Education production function and the equity of compulsory education," *Educational Research*, vol. 01, pp. 9-17, 2010.
[4] S. R. Sirin "Socioeconomic status and academic achievement: A meta-analytic review of research," *Review of Educational Research*, 2005, vol. 75, no. 3, pp. 417-453.
[5] Y. Yang and J. Gustafsson, "Measuring socioeconomic status at individual and collective levels," *Educational Research and Evaluation*, vol. 10, no. 3, pp. 259-288, 2004.
[6] X. Feng, "The characteristics and application of income measurement methods in surveys," *Social Science Research*, no. 3, 2006.
[7] N. Zhao, J. Fan, B. Yang, and Y. Ma, "The effect of family socio-economic status on pupils' language and mathematics achievement abstract," *Education Science*, vol. 02, pp. 69-74, 2014.
[8] D. R. Entwisle and N. M. Astone, "Some practical guidelines for measuring youth's race/ethnicity and socioeconomic status," *Child Development*, vol. 65, no. 6, pp. 1521-1540, 1994.
[9] J. Yu and Z. Xia, "Research on the problem of gray income in China's income distribution system," *East China Economic Management*, vol. 12, pp. 37-39, 2007.
[10] X. Wang, "Gray income and government reform," *Shanghai Economy*, vol. 05, pp. 14-15, 2011.
[11] X. Pan and G. Zeng, "Analysis of the reasons and strategies of fast increasing price of the real estate in China based on grey income," *Journal of Shaoguan University*, no. 1, pp. 86-89, 2013.
[12] W. Wang, "Study on developing big data industry," *Economics and Management of Technology*, vol. 01, p. 118, 2015.
[13] Investment Advisor, *China Big Data Industry Investment Analysis and Forecast Report 2020 –2016*.

[14] L. Zhang, "Social value and strategic choice of big data" Central Party School, 2014.
[15] W. Yang, "File information collection under the bid data background based on the theme of web crawler," *Lantai World*, no. 7, p. 20, 2015.

Xuesheng Qian is currently a master of business student in Antai College of Economics and Management, Shanghai Jiaotong University, China. He has a bachelor of computer science from Fudan University, China. Since 2012, he works as CTO in Shanghai Psylife Consulting Co., Ltd, which is leading K-12 data research institute of China with one great data warehouse assembling over one millions student's data, and involving a large number of public policy formulation. His research interests are big-data application, IT systems construction and business intelligence.

Yifeng Xu is graduated from the School of Psychology and Cognitive Science, East China Normal University, China. And he received a master's degree. Now he works in Shanghai Psylife Consulting Co., Ltd as the position of data analysis. He is devoted to the application of modern data statistics and mining methods in the field of education.

Jing Zhang got her master's degree in Antai College of Economics and Management, Shanghai Jiaotong University, China, majoring in management science and technology. Recently, she works as a data analyst in Shanghai Psylife Consulting Co. Ltd.



Wei Zhao graduated from the Hubei University of Economics with a degree in economics in 2015, and his major is statistics. He is the second author of the paper entitled "Analysis of market demand and its influence factors of life insurance company investing in retirement community", which has been contained in the *Proceedings of CICIRM2014*. His graduation thesis was on web scraping and Chinese text analysis. He is now working as a data analyst in Psylife Consulting Co., Ltd, Shanghai, China, focusing on data mining and visualization.