Quality Assessment of Regional Primary English Education Based on Big Data: A Case Study of Gansu Province

Jialan Nan (stnvjn6@ucl.ac.uk)
University College London https://orcid.org/0000-0002-9041-851X
Fu Yu
High School Attached to NENU: High School Attached to Northeast Normal University
xinying Xu
Zhangye high school in Gansu
Ren Feng
Wenxian No.2 high school in Gansu
Wenhui Liang
No.3 high school in Gansu province, Wenxian city

Case study

Keywords: Big Data, English Education, English excellence degree, Data mining, Gansu province

Posted Date: January 28th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1114036/v1

Abstract

Introduction: Recently, big data attract great attention from industries, academia, and governments. Education big data from teaching and learning processes used to assess elementary education level is scarce in China.

Case description: Gansu province, located in the northwest China, is one of the most backward economy areas where the development of elementary English education falls far behind developed areas of the country. Taking Gansu province as a case study and English subject in senior high schools as the study object, we use score data of English academic proficiency test (2018-2020) from 322 senior high schools to explore spatial-temporal differences of English education level among those schools in order to promote the balanced development of regional education. First we briefly describe the generation of English big data from senior high schools, and then, we mine the big data to analyze the difference of the English education level using the anomaly method, the comparison method, the cluster analysis method, and hot point analysis method. English excellence degree is developed as an indicator to assess elementary English education level.

Discussion and Evalution: The results show that: (1) there exists obvious difference in English education level among the selected senior high schools: the highest average value of English excellence degree is 759 times the lowest, and the number of schools with negative anomalies accounts for more than 50%; (2) English excellence degree displays an obvious agglomeration characteristic in Lanzhou city, the provincial capital, that is, the hot spots (high English level) of English excellence degree are located in Lanzhou, while the cold spots (low English level) are concentrated in minority and poor areas such as Linxia autonomous region, the major part of Longnan and the eastern part of Qingyang; (3) The investigation in the form of questionnaires confirms the fact that disadvantage schools in English level are usually lack of good teachers who have a preference to teach in areas with good conditions.

Conclusion: From the results, we can draw the conclusion that big data play an important role in quantitatively assessing basic education level, which lays foundation for improving the regional education level. Individuals' value orientation of study, attitude towards study and the study behavior will be investigated in the future.

Introduction

At present, there is no consensus on the definition of big data. It is generally accepted that it refers to data sets that are difficult to obtain, manage and process using common software tools within tolerable time slot. Big data is usually characterized by its huge volume in size, great variety in data types and velocity in fast data creating, processing and analyzing [3, 28]. Li & Cheng think that big data is as an important strategic resource as natural resources and human resources [21]. The US government believes that big data is the "new oil of the future" [7], so it takes the lead in launching the "Big Data Development Plan" to

make breakthroughs in multiple research and application fields, aiming at generating inestimable amount of value [33].

In terms of education, big data will promote the reform of teaching, educational research, educational management and educational evaluation [16, 37]. Especially in educational evaluation, big data can make it more comprehensive, objective and accurate [34]. Accurate assessment results let students know how well they have grasped the knowledge, understand their personality preference, make up for shortcomings and seek help on the nose. The assessment also lets teachers find out which knowledge has not been mastered by students and thus needs to be repeated or emphasized, and what kinds of problems are most prone to happen, so as to improve teaching methods. Moreover, it lets researchers explore the correlation between variables—learning content, learning time and learning method—and the final learning result [1, 22] in order to find main factors impeding student learning. Last but not the least, it lets policy makers realize the imbalanced development in education environment, education resources, education opportunities and education quality of regional education, and devote to establish a comprehensive, real-time and dynamic quality monitoring system of education [32]. It can be seen that the evaluation based on big data plays a huge role in the field of education. However, for the problem of serious imbalance of educational resources between urban and rural areas, among regions and across schools, there is still a lack of study on how we can creatively use big data to accurately evaluate the education quality so as to promote the balanced development of regional education.

Gansu is located in the western region of China, whose educational environment, especially that of the poor areas and minority areas, is significantly different from that of other regions. However, the contrast is only an empirical judgment. How to quantify the contrast is beyond the grasp of education policy-making departments. If there is no scientific, reasonable and systematic evaluation of teaching quality, but only the use of empirical generalization and subjective estimation to evaluate the difference in education, mistakes will be made in policy making. Through mining, analyzing, modeling and evaluating big data in education, managers can timely and accurately understand the regional teaching situation, find typical vulnerable schools and groups, and investigate the causes of the vulnerability, so as to more scientifically balance the allocation of education resources. English is an important subject in basic education in China [36]. Taking English education as an example, we aim to achieve the following objectives through the application of big data: (1) Scientifically reflect the spatial and temporal differences of regional English teaching quality; (2) Find some typical vulnerable areas (cold spots in English education) through big data mining. The research results provide methods for the education department to evaluate the regional education level, and help to make better education decisions in order to promote the balanced development of basic education.

Case Description

Study area

Gansu province is located at the intersection of the Loess Plateau, Qinghai-Tibet Plateau and Inner Mongolia Plateau, between 32°11'-42°57'N and 92°13'-108°46'E. Administratively, it has 14 cities including 86 counties. As of 2021, the permanent resident population is 25.0198 million, and the regional Gross domestic product (GDP) is 824.61 billion yuan, or 31,336 yuan per capita based on resident population [11]. Out of the 56 ethnic groups in China, there are 16 in Gansu, all of which have a population of more than 1,000. Due to multiple reasons, Gansu has been a relatively backward area in the development of basic education in China.

According to the statistics of Gansu Provincial Department of Education in 2019 [10], there were 376 senior high schools in the province (Figure 1), with 526270 students and an average population of 1399 in each school. As can be seen from Figure 1, the distribution of senior high schools in Gansu is extremely uneven, scattered in the western and southwestern part, and concentrated in the central and southeastern part.

Data collection

English education big data is composed of four parts: historical score data, basic information data, teaching data, and learner personal data (see Figure 2). Historical score data records the performance of the students generated in the learning process. Basic information data records basic information of schools, teachers, and students, such as the location of school, its education resource, teachers' and students' archive, etc. Teaching data includes the information of teaching materials, tools (internet) and teaching activities. Learner personal data contains information of students' individual character, preference and habits of study. In this study, we mainly use the historical score data for analysis.

Learning data varies spatiotemporally. Spatially, learning processes differ among schools in the region. Temporally, the scores fluctuate in mid-term, final and periodic test. We only use basic information data concerning school location points of interest (POI) (data source: http://www.databox.store/) and historical score data of English academic proficiency test (2018-2020) coming from Gansu Education Examination Institute for evaluating the English education level. Some data is collected from questionnaires.

Methodology

Through a variety of methods of data mining, we can identify some typical vulnerable schools, and then, design programs in line with the school's own development, such as teacher training programs and short-term exchange programs with high-quality schools, adjust the allocation of education funds, and implement a number of policies that favor vulnerable schools. The data mining methods and the paths to improve education quality are shown in Figure 3. In the next section, we introduce the analysis methods used to evaluate the English education level in the study area.

The anomaly method

The anomaly method is used to reflect the degree to which the performance data are away from the central value. Some typical vulnerable schools and vulnerable individuals can be distinguished according to this degree, which is represented by the anomaly percentage (k'), calculated as follows:

$$k' = \frac{k_i - \bar{k}_i}{\bar{k}_i} \times 100$$

1

where, k_i is an exam score in school i, $\bar{k}_i$ is the average of the exam score in all schools.

Comparison method

In order to assess the level of English in each school, we developed an index named English excellence degree (EED). The EED (X) is calculated as follows:

$$X_j = M_j \times E_j$$

2

where: M_j is the average score of English in the j school; E_j is the passing rate of English subject in the j school.

According to the EED from 2018 to 2020, we use the percentage change rate (in every two years) as the index to compare and analyze the changes in English teaching quality in each school, and look for schools with great changes.

$$\Delta EED = \frac{EED_t - \bar{EED}_{t-1}}{\bar{EED}_{t-1}} \times 100$$

3

where: ΔEED is the percentage change rate between year t and year t-1.

We use the quartile method to compare and analyze the changing trend of English quality. In this method, we arrange data from the smallest to the largest in each year. The first quartile is defined as the middle number between the smallest number (minimum) and the median of the data set. The second quartile is the median of a data set. The third quartile is the middle value between the median and the highest value of the data set. We can use those quartile values to build a violin plot. A violin plot is a method for graphically depicting groups of numerical data through their quartiles. Moreover, it also shows the probability density of the data at different values. We use the GraphPad Prism 8.0 software to make the

violin plot, which by reflecting the distribution characteristics of EED every year, helps us compare the English level of all schools in three years.

Cluster analysis method

Clustering is the process of grouping data into groups or clusters, so that data within the same group retain high level of similarity but groups differ from each other on measuring index. Here, similarities and differences are assessed based on the EED of schools, which often involves distance measures [35]. We use the Euclidean distance method to determine how different EED are among schools in this study. There are *n* clusters (here referring to schools), d_{ij} represents the distance of EED between school I_i and school I_j which can be calculated by the following formula:

$$d_{ij} = \sqrt{\left(I_i - I_j\right)^2}$$

4

After d_{ij} is calculated, a distance matrix $D = \left(d_{ij}\right)_{n \times n}$ is obtained. Given the distance matrix D, the procedure for grouping *n* schools proceeds in the following steps. Merge a pair of schools whose d_{ij} is the smallest, leaving *n*-1 clusters for the next step. Next, repeat step one on (n-1) clusters, and get *n*-2 clusters in total. Then, continue doing this procedure—reducing one cluster at each step—until only one cluster, containing *n* schools, is formed. The groupings process is shown in the tree diagram (the dendrogram).

Hot spot analysis method

Hot spot analysis based on Kernel density estimation (KDE)

Kernel density method is an important statistical analysis method for extracting the distribution characteristics of geospatial attributes. It is based on the first law of geography, namely, the law of distance attenuation effect. The closer the things are, the more correlated they are to each other. The Kernel density is calculated by the quadratic kernel function $\hat{f}(x, y)$[36], whose expression is:

$$\hat{f}(x, y) = \frac{3}{nh^2\pi} \sum_{i=1}^{n} \left[1 - \frac{\left(x - x_i\right)^2 + \left(y - y_i\right)^2}{h^2}\right]^2$$

5

where, x_i and y_i are the coordinates of grid *i*, n is the total number of schools, *h* is the bandwidth parameter, that is, the distance attenuation threshold. $(x\text{-}x_i)^2 + (y\text{-}y_i)^2$ represents, within the range of *h*, the square of the Euclidean distance between grid *i* and sample point (*x*, *y*). The selection of *h* in the Kernel density formula is often determined by the dispersion degree of point of interest (POI) and the analysis scale. 5000m is selected according to the size of the study area and the distribution of schools. First, the

study area is grided, the density contribution of each school to the center point of each grid in the range *h* can be calculated using the Kernel function (f(x,y)). With multiple schools, a density map is generated by spatial superposition of the density values of each grid cell. The area with high density value is a hot spot, otherwise, it is a cold spot area.

GIS based spatial hot spot detection and analysis method (Getis ord G_i)

GIS based spatial hot spot detection analysis is a method of spatial autocorrelation at the local scales [32] based on distance weight. It is used to measure the correlation, which is represented by the G_i index, between an attribute value at a location and the same attribute value at a neighboring location (here the attribute value is represented by EED). When G_i is high, it indicates the presence of a hot spot. When G_i is low, it indicates a cold spot instead. The G_i index is calculated as follows:

$$G_i = \frac{\sum_j^n W_{ij} x_j}{\sum_j^n x_j} \text{ (j≠i) (6)}$$

where: x_j is the attribute value (EED) of school *j*; *n* is the total number of schools; W_{ij} is the spatial weight between school *i* and school *j* within distance *d*. G_i can be standardized in this way:

$$Z\left(G_i\right) = \frac{G_i - E\left(G_i\right)}{\sqrt{VAR\left(G_i\right)}}$$

7

where $E(G_i)$ is the mathematical expected value and $VAR(G_i)$ is the coefficient of variation. When $Z(Gi)$ is positive, it indicates that the value around the spatial unit *i* is relatively large, which means the spatial unit is a hot spot area; When $Z(Gi)$ is negative, it indicates that the value around the spatial unit *i* is relatively small, which means the spatial unit is a cold spot area. A fishnet with a 5000×5000m grid is created with Create Fishnet tool in the ArcGIS, and the then it is cut by the boundary of Gansu province. After that, Zonal Statistics as Table tool and Join tool are used to link the distribution data of schools with EED attributes to fishnet files. Finally, Hot Spot Analysis with Rendering is used to obtain the hotspot distribution map.

Discussion and Evaluation

Evaluation

Analysis of differences in English education level

A total of 322 senior high schools in the province are selected. Their average EED in the three years ranges from 0.21 to 91.12. Schools that have EED above the average account for 43.37% of all schools,

and those that have EED below the average account for 56.63%. In 2018, the schools with positive anomalies account for 43.48% of the total, and those with negative anomalies account for 56.52%. The highest anomaly percentage is +188.89%, and the lowest anomaly percentage is -98.55%. In 2019, there are 42.55% of the schools with positive anomalies and 57.45% with negative anomalies. The highest anomaly percentage was +185.53% and the lowest was -99.47%. In 2020, 44.10% of the schools are with positive anomalies and 55.90% are with negative anomalies. The highest percentage anomaly is +165.69%, while the lowest is -100% (Fig. 4). The data of three years of anomaly shows that English performance varies greatly. The lower the anomaly value is, the more disadvantaed the school is. According to our analysis, 71 schools have anomaly values below -70% and 32 schools have anomaly values below -90%, which indicates that they are poor in terms of English education quality.

The schools from cluster analysis are divided into four categories. There are 33 schools in category four, whose EED values range from 71.90 to 91.85, with the mean being 80.22. Category three includes 76 schools that EED of these schools vary from 46.49 to 69.92 with the mean value being 57.62. In category two, there are 74 schools. EED in these schools are from 27.07 to 45.14, with the average being 35.16. Schools in category one are 139 whose EED range from 0.27 to 26.16, with the mean being 10.81. From cluster analysis, we see that many “poverty” schools in English education (Fig. 5) that distribute in the southeast part of Gansu province.

Analysis on the interannual change of English education level

There is a significant difference between the percentage change of EED in 2018-2019 and that in 2019-2020 (see Fig. 6). From year 2018 to 2020, 186 schools have improved their English teaching quality, accounting for 57.76% of the total number of schools, while 136 schools have decreased their English teaching quality, accounting for 42.24% of the total number of schools. We regard year 2018-2019 and year 2019-2020 as two stages, calculate the change rate of each stage, and then the change rate of two stages. The schools whose absolute change rate of two stages is more than 100 are shown in Table 1. The school that has declined the most in English teaching level is No. 321, and that has improved the most is No. 293.

Table 1
Change of English level from 2018 to 2020

The order of change rate	No. of school	Change rate between 2018 and 2019(%)	Change rate between 2019 and 2020(%)	Change rate of two stages(%)
1	321	288.94	-85.61	-374.55
2	306	241.16	-36.51	-277.67
3	84	245.66	-19.70	-265.36
4	227	205.66	-28.26	-233.92
5	186	188.11	9.89	-178.22
6	316	-35.48	117.12	152.60
7	253	19.48	173.46	153.98
8	49	6.63	163.45	156.82
9	107	-53.21	104.26	157.47
10	196	-67.08	95.63	162.70
11	46	-29.76	141.12	170.88
12	109	-49.28	128.06	177.34
13	215	32.27	213.31	181.05
14	189	-56.69	142.62	199.31
15	288	-36.47	170.83	207.30
16	310	-57.47	198.14	255.60
17	117	-67.73	188.82	256.56
18	287	-54.37	202.30	256.68
19	86	-41.54	229.44	270.98
20	308	-50.53	246.55	297.08
21	114	-65.40	233.82	299.23
22	226	-69.04	230.61	299.65
23	188	-38.88	269.66	308.54
24	120	-42.45	284.78	327.23
25	108	-61.93	319.64	381.57
26	293	-75.86	331.37	407.23

We use three violin plots to show the distribution of EED for all schools from 2018 to 2020 below (See Figure 7). We can see that the plot of 2018 is similar to that of 2019. The plot of 2020 looks different in that its EED values shift up, indicating an overall improvement in English teaching level.

Regional Differences in English education level

We use Kernel density tool in ArcGIS 10.6 software to analyze the spatial distribution of Kernel density in the study area (Fig. 8). The results show that the density is highest in Lanzhou, the provincial capital, and second highest in Tianshui, Pingliang, Wuwei and Jiuquan. High-density areas represent hot spots of EED, which are regions with high English education quality.

With the help of the spatial hot spot detection and analysis tool in ArcGIS10.2 software, we obtain the distribution of Z(Gi) values.Hot spot and cold spot regions are indicated by Z(Gi) (Figure 7). If Z(Gi)>2, the region is a hot spot region with high English teaching level; if Z(Gi)<-2, the region is a cold spot region with low English teaching level. As can be seen from Figure 9, the hot spots are mainly concentrated in Lanzhou and the central part of Hexi Corridor, while the cold spots are concentrated in the southern Gansu, mainly in Linxia city, the major part of Longnan and the eastern part of Qingyang.

Discussion

Differences in English education level in the study area

From the English proficiency test results, English education level varies greatly among different schools. In 322 senior high schools, the highest average value of EED is 759 times the lowest. Schools with EED below the average value account for more than half of the total number of schools (56.63%) in 2018-2021. 71 schools have anomaly values below -70% and 32 schools have anomaly values below -90%, which indicates that they are poor in terms of English education quality. We carried out investigations on those 32 schools and found that most of them are located in remote mountainous areas and the residental areas of ethnic minority groups. Disparities in educational achievements resulted from ethnic and racial factors are a core issue of educational research. English as a global language is widely used in the world [31], and it is one of the compulsory courses in primary education in China [37]. Students in ethnic areas learn not only their mother tongue, but also English and Chinese as well. Therefore, they encounter more challenges when learning English. First, ethnologue and English belong to different language categories, having different phonetics and grammar characteristics. Second, the learning environment of students in ethnic areas is poor---with a lack of available teaching resources and professionally well-trained English teachers who, if do work there, are asked to speak and understand the ethnologue [6, 9]. Third, there exists negative attitude of the schools and patents toward English learning. These challenges make students in ethnic areas learn in a helpless and anxious way, and such mindset also impact the learning performance of subjects other than foreign language [16, 28, 41]. Some researchers think that teachers and their teaching practice are prominent factors in exacerbating or alleviating learned helplessness symptoms and behavior [4, 14]. Some suggest that negative attitude or

no involvement of parent results in poor academic performance [21, 24]. In summary, it is difficult for students in ethnic areas to learn English. In order to resolve the problem and to eliminate the big differences in English education level in the study area, Chinese State Education Commission has released many educational policies such as increasing salaries, implementing a rotation system, giving priority to professional title promotion and so on [19].

The level of English education changes significantly from year to year

From the analysis of the percentage change of EED in three years, we find that the English level of many schools vary greatly from year to year, with some increaing and others decreasing. However, if we compare stage 2018-2019 to stage 2019-2020, we can see that the overall English level has improved in the latter one.

The influencing factors of English performance include internal factors and external factors. The most basic and indispensable factors are teachers and students, while other factors exert their influence on Enlish performance via teachers and students [26]. Previous studies show that the allocation of teacher resources has an important impact on the examination results [20]. Duflo et al. find out that reducing the ratio of student to teacher can improve the quality of education by reducing class size and hiring more local teachers [8]. Blake et al. point out that the racial and ethnic status of students and teachers would affect students' educational level [5]. Generally, teacher resources keep stable in high-quality schools, while low-quality schools have a higher turnover rate of teachers than high-quality schools, and private schools have a higher turnover rate than state schools [2]. The results from questionnaries in the study is the consistent with that in literature[2], in addition, the teachers who quit are relatively younger and have shorter teaching experience compared to before. Teacher turnover is disruptive for students' academic attainment [15]. Teaching environment, salary and welfare guarantee are responsible for teacher turnover and loss [25]. Besides, the principal shortage and weak leadership are other contributors to the loss of teachers [13], whose impact is even more severe in schools with high concentrations of ethnic minority students and poor students [30]. The educational regulation has an important influence on the improvement of teacher allocation in basic education in Gansu Province. Local governments have increased funding for education, improved educational facilities and increased the spots for teachers in ethnic and poor areas. The intensity of improvement was the greatest in 2018 [10]. Corresponding improvement of English academic test scores might happen during 2019-2020.

Aggregation pattern in English education level

From our analysis of the hot spot, we see that the regional differences in English level are significant. The English education quality decreases as we move away from the provincial capital (Lanzhou) to the periphery cities. On the provincial level, the regions with high English level are concentrated in Lanzhou and the central part of Hexi Corridor, while the regions with low English level are mainly concentrated in Linxia city, most parts of Longnan and eastern part of Qingyang. This shows an aggregation pattern. The spatial inequality of education can be between provinces, between rural and urban areas, or between districts or counties [12, 27]. Peng points out the spatial difference is mainly due to unbalanced economic

development [33]. Many researchers focus on the factors affecting the disparties. Yu et al. point out the regional differences mainly result from the difference of education resources allocation [42]. Hussar and Sonnenberg find out the main factor is the difference of expenditures per Pupil [18]. Kang et al. summarize that the regional differecne of education is influenced by many factors such as geographical location, educational fund, and the quality of teachers and students [20]. Considering many factors mentioned above, the difference of education of inter-cities is greater than that of intra-cities [42]. Just because of the inequality in education, the selection of school has become an inevitable problem for both the students and the teachers in the stage of compulsory education, which makes the inequality even larger. Equal opportunity to education is a basic human right championed by the United Nations [39]. In order to balance the development of basic education, effective countermeasures have been put forward by Chinese goverment, for example, balancing allocation of basic educational resources, increasing the investment in weaken areas and schools, encouraging teachers for training, achieving teacher shift, et al. Shanghai has taken the measures of collectivization in running a school. At the micro-management level, for collectivized schools, their school management structure will be optimized, professional development of teachers enhanced, and all-round development of students promoted [40].

Conclusion

According to the results obtained in the study, it is concluded that there seems an obvious regional difference in English education level. Disadvantage schools are due to a lack of high-quality teachers. Allocation of teacher resources is a main factor which influences student English performance. We see that some of the research results are in support of the conclusion derived from this study. Many educational policies made by Chinese State Education Commission and local goverment should favor poor or minority areas.

Abbreviations

EED: English excellence degree; GDP: Gross domestic product; US: United States; POI: Points of interest; KDE:Kernel density estimation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Availability of data and materials

The school location points of interest (POI) used in the study is available at the data portal (http://www.databox.store/). The historical score data of English academic proficiency test (2018-2020) are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the key project of Gansu Provincial Education Planning from Education Department of Gansu Province (GS[2020]GHBZ214).

Authors' contributions

The corresponding author is the sole owner and producer of all the research methods and writs the original draft,Fu Yu isthe funding gainer. Xinying Xu takes part in questionnaire survey in Zhangye city, Feng Ren takes part inquestionnaire survey in Wenxian county, and Wenhui Liang takes part inquestionnaire survey in Wenxian county.All authors readand approved the final manuscript.

Acknowledgements

Not applicable

References

1. Aleksandra KM, Ivanovi M, Budim AZ(2017)Data science in education: Big data and learning analytics. Computer Applications in Engineering Education 25(6):1066-1078.
2. American Institutes for Research(2011)http://www.air.org/project/schools-and-staffing-survey-sass. Accessed 30 Nov 2011.
3. Bai X, Zhang F, Li J (2021). Educational big data: predictions, applications and challenges. Big Data Research 26:1-12.
4. Biber M, Biber, SK (2014) Investigation of the level of prospective teachers' learned helplessness in mathematics in relation of various variables. Procedia Social & Behavioral Sciences 116: 3484-3488.
5. Blake JJ, Smith DM, Marchbanks MP(eds) (2016) Doesstudent-teacher racial/ethnic match impact blackstudents' discipline risk? atest of the cultural synchronyhypothesis. Palgrave MacmillanUS,New York.
6. Brown JS, Collins A, Duguid P(1989) Situated cognition and the culture of learning. Educational Researcher.(1):32-41.
7. Cheng XQ, Jin XL, Wang YZ (2014) Survey on big data system and analytic technology. Journal of Software 25(9): 1240-1252.

8. DufloE, Dupas P, KremrM (2015) School governance, teacher incentives, and pupil-teacher ratios: experimentalevidence from Kenyan primary schools. Journal ofPublic Economics123:92-110.
9. Feng Zhiwen, Yuan Yichuan (2020) A study of basic foreign language education in ethnic areas of Yunnan in the context of language-based poverty alleviation. Journal of Yunnan Normal University (Humanities and Social Sciences Edition) 52(5):31-40.
10. Gansu Education Department office, Gansu Institute of Educational Sciences (eds) (2018). Gansu education yearbook. Gansu Education Press,Lanzhou China.
11. Gansu Province Bureau of Statistics (ed) (2020). Gansu development yearbook. China Statistics Publishing House,Beijing.
12. Gao Y, He Q, Liu Y, Zhang L, Wang H, Cai E (2016) Imbalance in spatial accessibility to primary and secondary schools in china: guidance for education sustainability. Sustainability8(12):1236.
13. GatesSM,Jeanne SR, Lucrecia S, Catherine HC,Karen ER(eds) (2003) Who is leading our schools? an overview of school administrators and their careers. RANDCorporation,SantaMonica, CA.
14. Ghasemi F (2021)A motivational response to the inefficiency of teachers' practices towards students with learned helplessness. Learning and Motivation.doi:10.1016/j.lmot.2020.101705.
15. GibbonsS, Scrutinio V, Telhaj S (2021) Teacher turnover: effects, mechanisms and organizational responses. Labour Economics.73: 102079.
16. Horwitz E (2001) Language anxiety and achievement. Annual Review of Applied Linguistics 21:112-126.
17. Hu Bicheng, Wang Zulin (2015) The role, challenge and trend of education reform of big data: a review of the latest research progress of education reform in the era of big data. Modern University Education (4): 98-104.
18. Hussar W, Sonnenberg W (2000)Trends in disparities in school district level expenditures per pupil. Education Statistics Quarterly2(1):74-75.
19. Jia Y, Jiang F (2015). The enlightenment of the balanced development strategy of basic education in China and abroad. Education Science Forum (7):78-80.
20. Kang W, Li D, Liu H (2021) The realistic dilemma and transcendental path of balanced development of compulsory education. Education Science Forum(12):67-71.
21. Lee J, Barror J (2001) Schooling quality in across-sectionof countries. Economica68:465-488.
22. Li Guojie, Cheng Xueqi (2012) Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences 27(6):647-657.
23. Li Y, Zhai X (2018) Review and prospect of modern education using big data. Procedia Computer Science 129:341-347.
24. LiuG, Teng X (2012) The relationship between parental involvement and immigrant children's academic achievement: the mediating role of autonomous motivation. Psychological Exploration36(5): 433-438.

25. Liu S, Onwuegbuzie AJ (2012) Chinese teachers'work stress and their turnover intention. International Journal of Educational Research 53:160-170.
26. Luo Junbing (2018) Research on influencing factors of basic education quality in ethnic areas of China. Basic Education Research (9):24-27.
27. Ma G, Wu Q (2019) Social capital and educational inequality of migrant children in contemporary China: a multilevel mediation analysis. Children and Youth Services Review 99:165–171.
28. MacIntyre PD, Gardner RC (1994) The subtle effects of language anxiety on cognitive processing in the second language. Language Learning 44(2):283-305.
29. Manyika J, Chui M, Brown B (eds) (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, Washington, DC.
30. McKibbenS (2013)Do local-level principal preparation programsprevent principal turnover? evidence from the 2008-2009 schools and staffing survey (SASS)principal follow-up survey. The Public Purpose(11): 69-87.
31. Melitz J (2016) English as a global language. In: GinsburghV, WeberS (eds) The palgrave handbook of economics and language. Palgrave Macmillan,Houndmills, UK.
32. Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. Geogr. Anal.27(4): 286-306.
33. Peng H, Qi L, Wan G (2020). Child population, economic development and regional inequality of education resources in China. Children and Youth Services Review. doi:10.1016/j.childyouth.2020.104819.
34. Pulse UG (2012) Big data for development: opportunities &challenges. http://www.unglobalpulse.org/sites/default/files/Big-Data for Development-UN Global Pulse June 2012.pdf. Accessed 21 June2012.
35. Ramos TG, Machado J, Cordeiro B (2015) Primary education evaluation in Brazil using big data and cluster analysis. Procedia Computer Science55:1031-1039.
36. Silverman BW (ed) (1986) Density estimation for statistics and data analysis.Chapman and Hall,New York.
37. Ministry of Education of the People's Republic of China (MOE) (ed) (2017) The curriculum scheme for the senior high school. People's Education Press, Beijing.
38. S Department of education. Enhancing teaching and learning through educational data mining and learning analytics]. http://www.ed.gov/edblogs/technology/files/2012/03/edm-la-brief.pdf.Accessed 12 Oct. 2012.
39. World Education Forum Drafting Committee (2000) Education for all: meeting our collective commitments. Notes on the Dakar framework for action.http://www.worldfamilyorganization.org. Accessed 23May 2000.
40. Wang Yutian (2020) The practical research on group education in the area from the perspective of the balanced development of compulsory education—taking J district of Shanghai as an example.

Dissertation, East China Normal University.

41. Wen Qingxia (2014)A review of learned helplessness. Journal of Jiangsu University of Technology20(1): 64-70.
42. Yu Yang, Han Zenglin, Peng Fei, Liu Tianbao (2016)Spatio-temporal changes of the compulsory education resources allocation difference in Liaoning province 35(6):21-26.

Figures

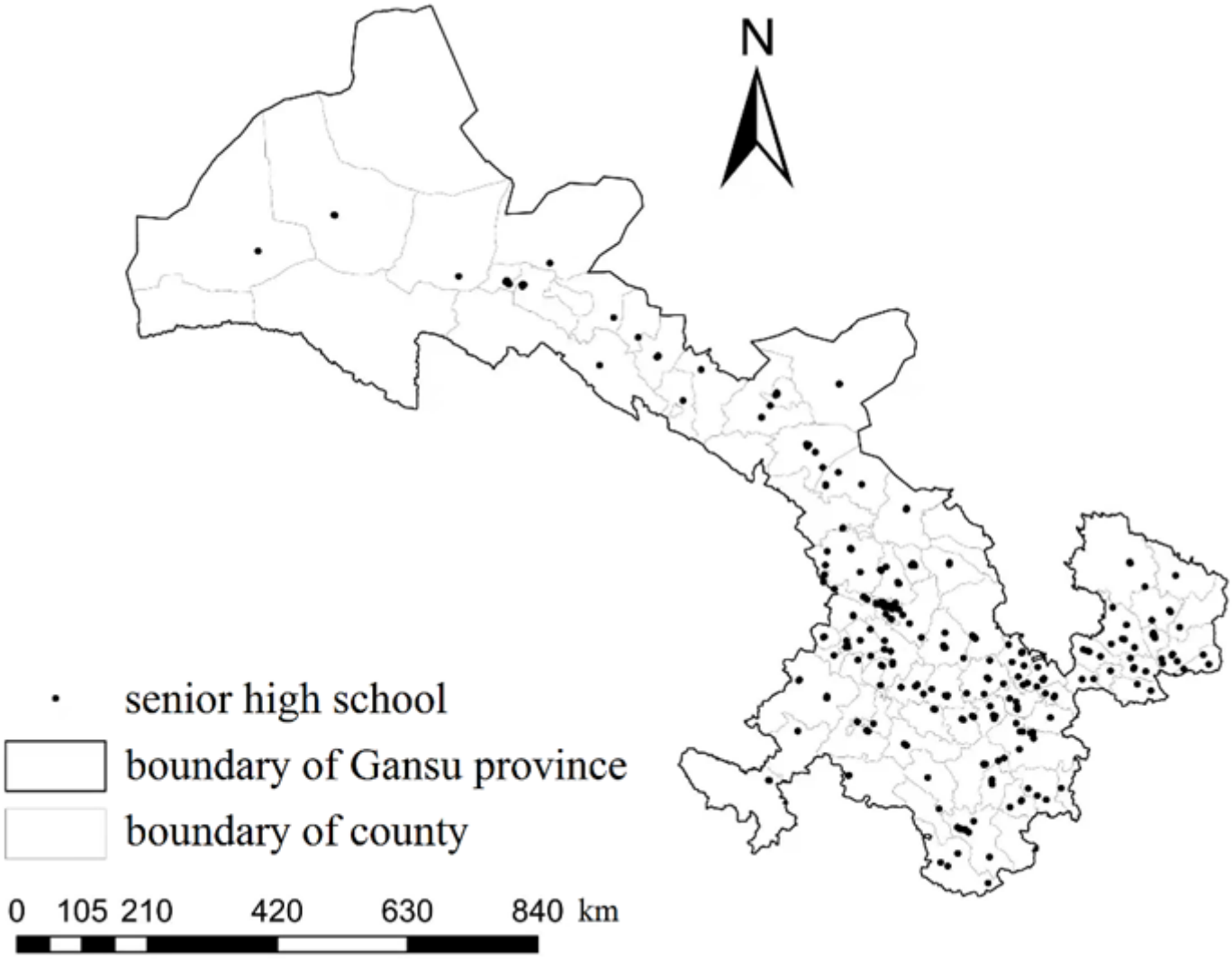

Figure 1

Distribution of senior high schools in Gansu province

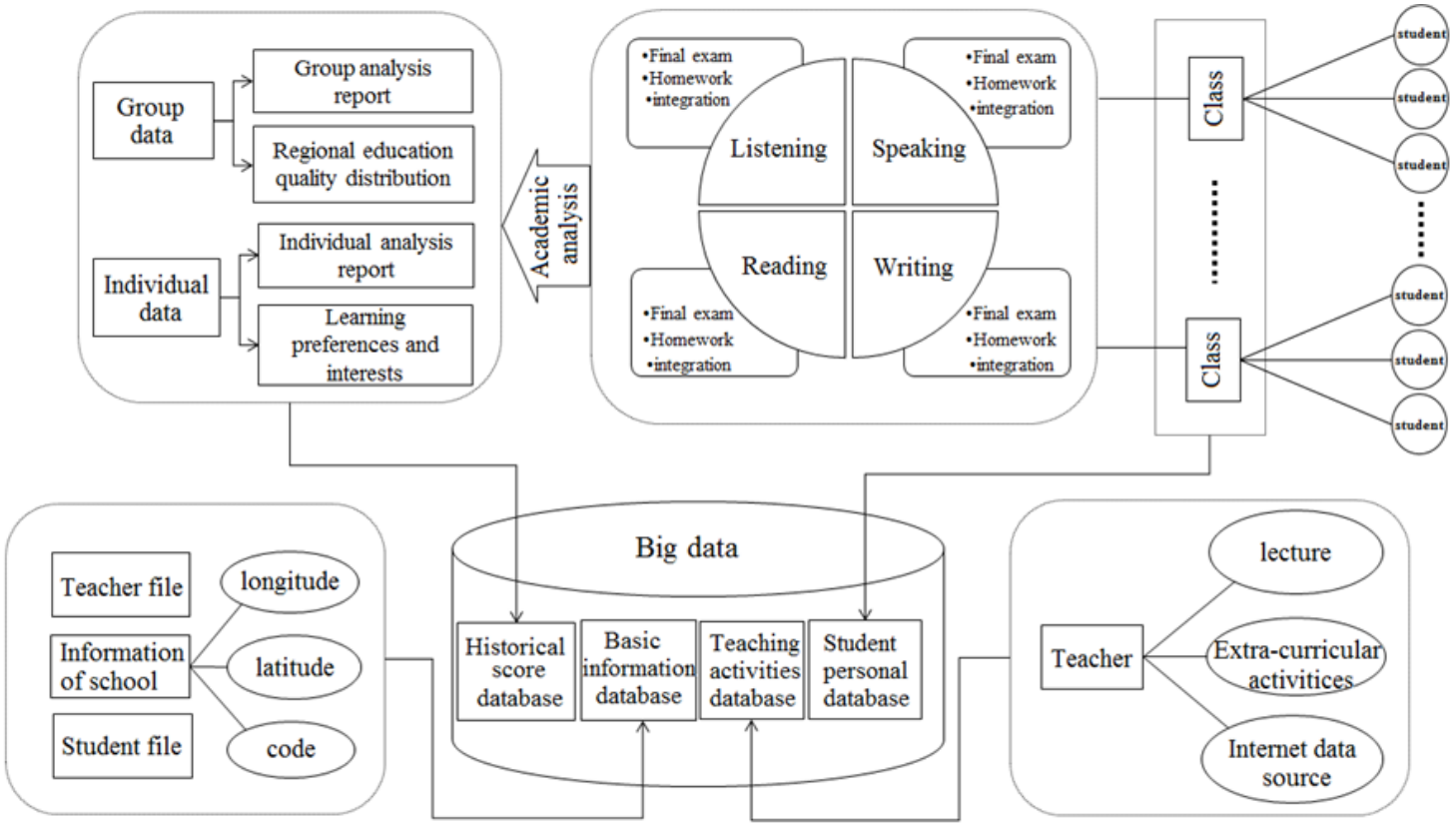

Figure 2

Framework of English educational big data generation

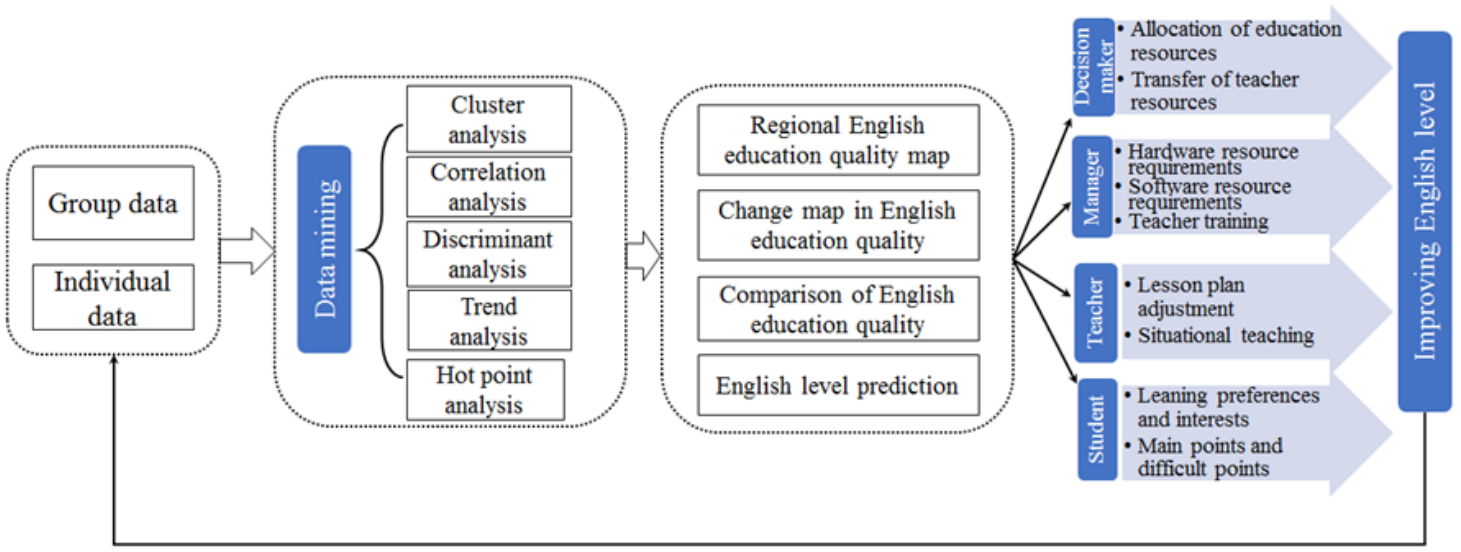

Figure 3

Data mining and the framework of analysis system

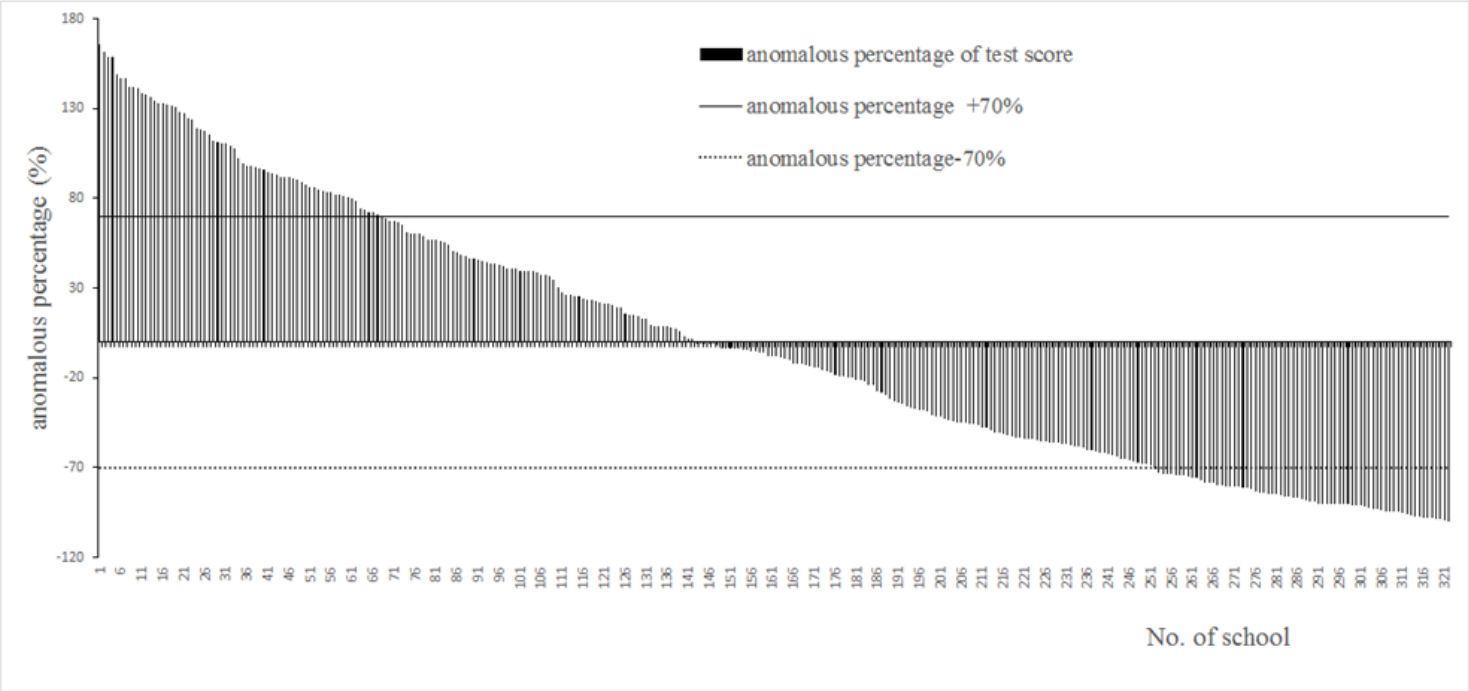

Figure 4

Distribution of anomalous percentage of English test score in 322 senior high schools in 2020

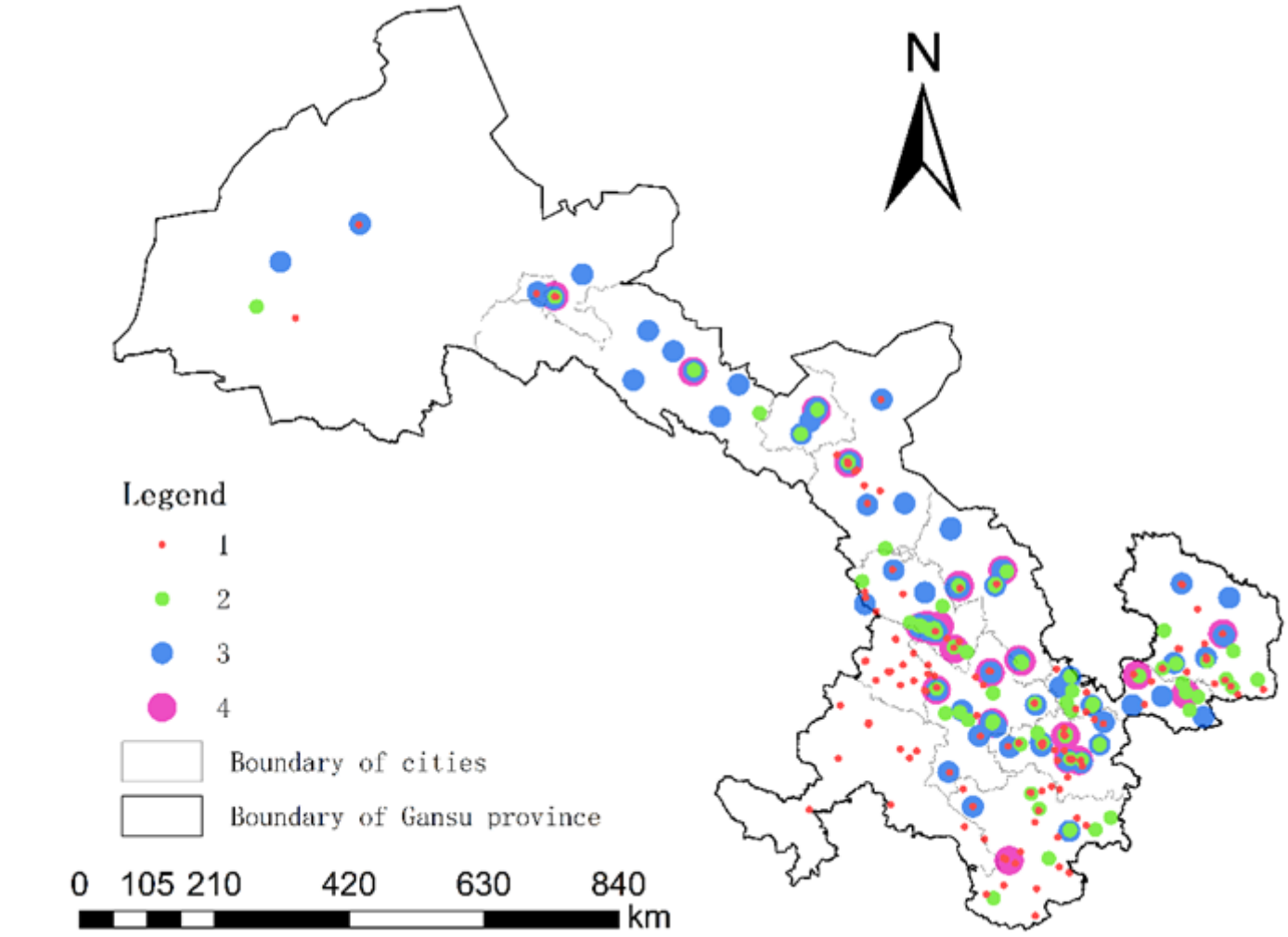

Figure 5

Distribution of four categories from cluster analysis in 322 senior high schools in 2020

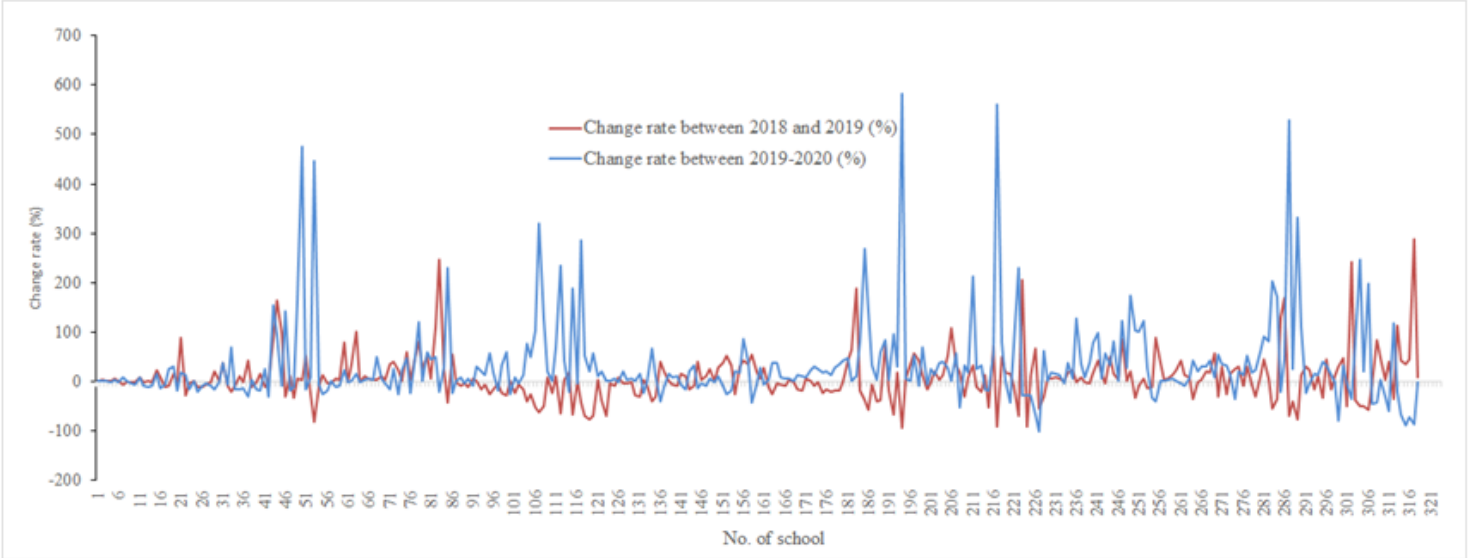

Figure 6

Comparison of English teaching level in the two stages in 322 senior high schools

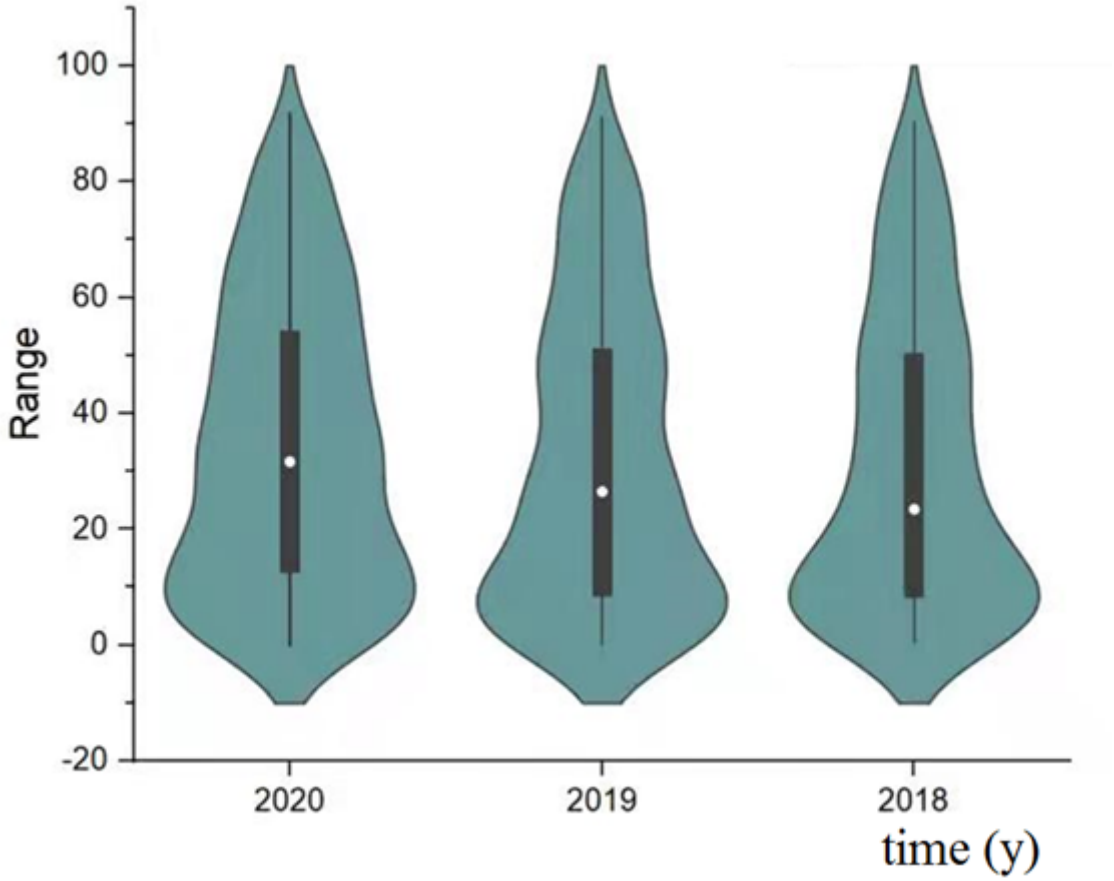

Figure 7

Annual differences in EED. The white dots present the median, the thick vertical bars represent the interquartile range, and the thin vertical lines extend to 1.5 times the interquartile range. The density plot width of the violin area presents the frequency distribution.

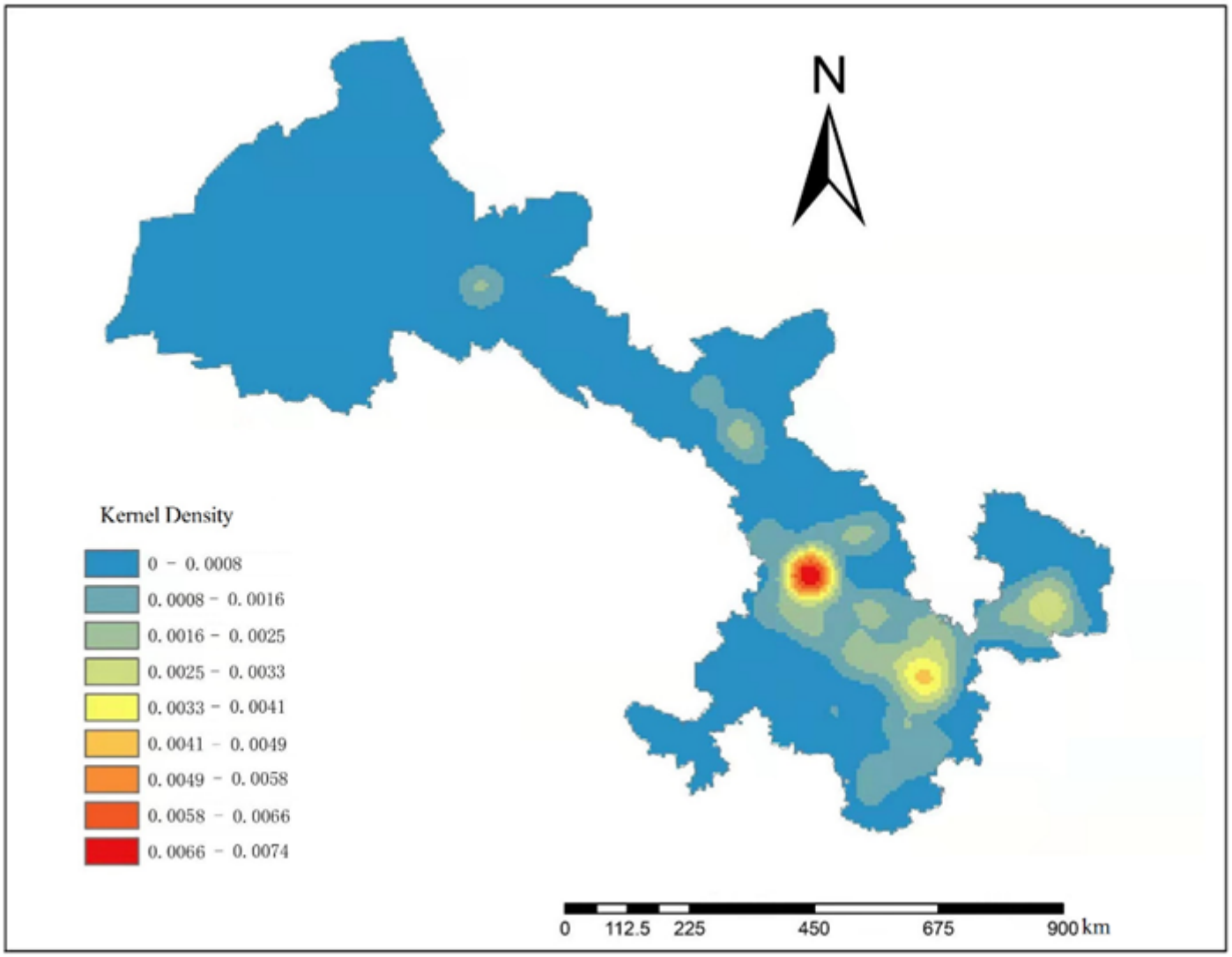

Figure 8

Spatial distribution of Kernel density of EED in the study area

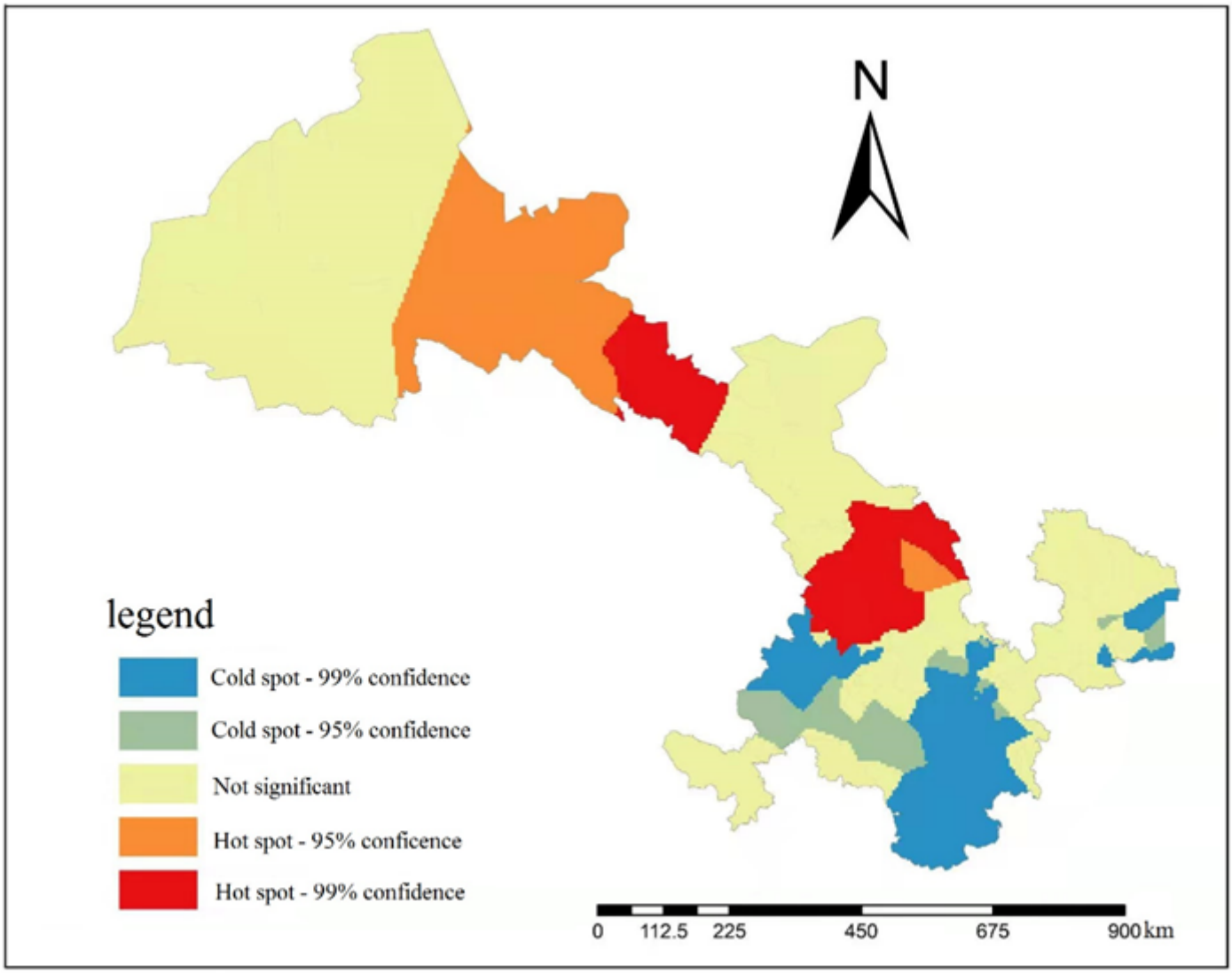

Figure 9

Hot spot maps showing of spatial clustering patterns for EED: red color represents hot spots; blue color expresses cold spots.