ICIC Express Letters
Part B: Applications
Volume **12**, Number **6**, June **2021**
ICIC International ©2021 ISSN 2185-2766
pp. 515–523

ESTIMATING STUDENT-TEACHER RATIO WITH ARIMA FOR PRIMARY EDUCATION IN FLUCTUATING ENROLLMENT

Chia-Chi Chen[1,*], Yujie Chen[2] and Cheng-Yi Kang[1]

[1]Doctoral Program of Educational Leadership and Technology Management
Tamkang University
No. 151, Yingzhuan Road, Tamsui District, New Taipei City 251301, Taiwan
[*]Corresponding author: sophiabv03@gmail.com; sophiachen@g2.usc.edu.tw
serev31826@gmail.com

[2]College of Education
Northeast Normal University
No. 5268, Renmin Street, Nanguan District, Changchun 130024, P. R. China
chenyj471@gmail.com

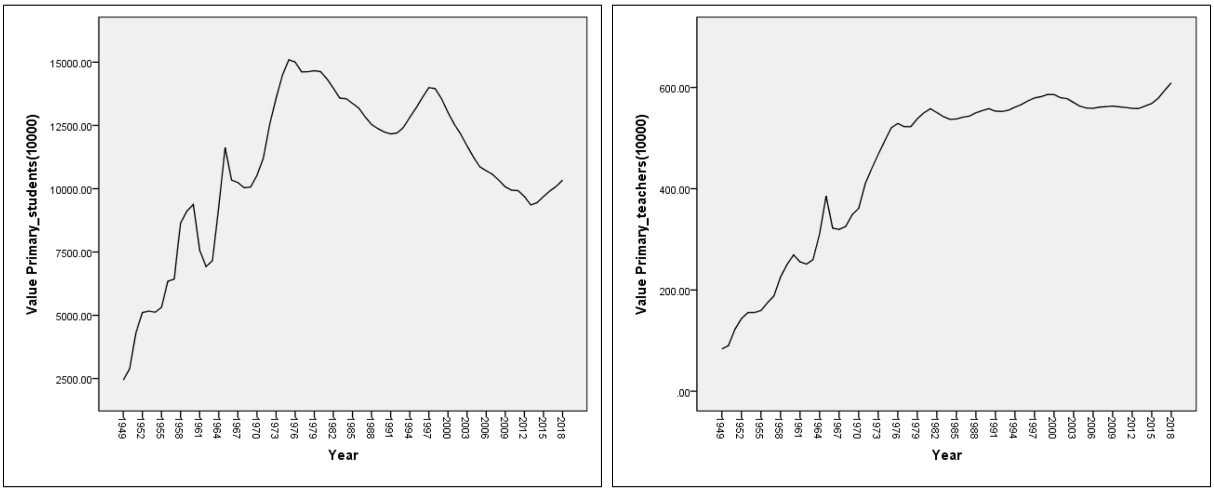
Received October 2020; accepted January 2021

Abstract. *This study develops a framework to detect the trend of student-teacher ratio (STR) in the fluctuating student enrollment in primary education. The target education series data (1949-2018) were cited from the MOE in China. Taking China's primary education as an example, we employ cross correlation function, ARIMA or ARIMAX model to verify the future trends of student and teacher numbers for STR. The findings suggest the fittest ARIMA models were used to interpret the trend of STR properly. The findings reveal the calculated STR with the trend of fluctuating student and teacher numbers will decline in future. While the predicted STR might increase in future based on the proposed ARIMA model with STR series data. The gap between the calculated and predicted STR may provide a feasible range to estimate the future trend. The results provide useful information for related policy makers to better control the quality of primary education.*

Keywords: ARIMA, Cross correlation function, Primary education, Student-teacher ratio, Teaching quality

1. **Introduction.** At the time of 1998, there are 600,000 primary schools in China; more than 139 million primary students enrolled with no diversification of curriculum having 5.8 million teachers in this primary education system [1]. According to the National Bureau of Statistics of China, the yearbook mentioned the transformation of primary education typically is driven by content, curriculum reform, social change, and even the declining birthrate. While the structure of student-teacher ratio (STR) issue in primary education under the pressure of declining birthrate has little been discussed with series data in previous literature [2]. Even though China has the largest primary education system in the world, the declining birthrate may show a potential risk threatening the expanding primary education. Based on the data from Ministry of Education in China, we found the largest number of student in primary schools is 150,941 thousands (1975), and recent lower is 93,605 thousands (2013) [3]. The declining of student numbers is amazing, see Figure 1(a). On the other hand, the number of teachers has shown increasing in China steadily. The largest numbers of teachers have shown in 2018, see Figure 1(b). We wonder the increasing of teachers has moved to a new stage and might grow unlimited in future. This study aims to realize to what extent the declining will impact the student and teacher numbers in elementary schools directly. We will focus on the quality of teaching issue in terms of the STR which will impact the quality of classroom teaching. Based on the long

DOI: 10.24507/icicelb.12.06.515



(a) (b)

Figure 1. The trend of student and teacher numbers in primary schools in China

term data set, we will analyze the STR in China and select fittest models to project its trend in future. In this sense, time series analysis could be an optimal choice in this study. Time series studies occur in the field of economics, where they exposed daily stock market quotations or monthly unemployment figures [4]. Various time series studies have focused on ARIMA (autoregressive integrated moving average) models with serial data [5-7], while using series data is an emerging area of research in social science [8-11]. Although time series analysis has been used for a long time, its application to tackle STR issue is still limited. This study may fill the research gap in current education studies.

The number of teachers per class and the STR are indicators of quality teachers in a country. The indicator may offer policy insights into opportunities for teacher preparation or the allocation of teachers properly. Regarding the global context, the fitted STR is diverse. OECD-related data may provide a guideline to review this issue. On average across OECD countries, class size decreased between 2005 and 2015 in 13 out of the 25 countries at primary education. Generally, at elementary school level, there are 21 pupils in an average class in OECD countries. While the OECD report reveals that there are fewer than 27 pupils per class in nearly all countries with the exception of Chile, China, Israel and Japan [12]. Furthermore, the OECD 2017 report shows the ratio of students to teaching staff in primary education is on average 15 in OECD countries, on average 14 in EU22 and 19 in G20 with full-time equivalents. The STR ranges from 10 or fewer in Lithuania and Norway to 27 in Mexico, 29 in India and 33 in South Africa [12]. The STR has become one of essential quality indicators in primary education. In this study, we will explore to what extent of the gap between China and OECD countries. Specifically, this study will explore the STR during the student enrollment declining in China. Given this purpose, this study will address the following research questions:

a) What kind of relationship between student and teacher numbers in primary schools in China?

b) Can STR be predicted by using time series approach properly?

c) What kind of trend of STR under decline birthrate is in the future?

The structure of this paper will be displayed as follows. First, we define the data set and their transformation process. Second, we display the result of ARIMA with student-teacher ratio prediction. Finally, the conclusions and suggestions are drawn.

2. **Method.** The related primary education series data (1949-2018), cited from the MOE in China, were published online or printed for public purposes [3]. The major technical terms in this study are defined as follows.

- Capacity of primary school students (CPSS) refers to total number of primary school students in the education system. The series data were collected from 1949 to 2018 based on annual basis.
- Capacity of primary school teachers (CPST) refers to the total teachers in primary schools on an annual basis. The series data were collected from 1949 to 2018. Totally, there are 70 records.
- Student-teacher ratio (STR) refers to numbers of students over teachers in primary education. Original student-teacher ratio is based on the report of MOE in China.

Dealing with forecasting, we found regression model, trend analysis and ARIMA model have different functions with data sets. The regression model typically fits randomized data sets; in contrast, time-series data is considered to its normalized assumptions. The trend analysis belongs to the time series family, and it is limited to handle one data set for each model only. The selected series data sets covered 70 periods to fit the requirement of ARIMA or ARIMAX (multivariable autoregressive integrated moving average) model building. Considering the characteristics of the data sets, we selected the ARIMA or ARIMAX model to build fitted future trends. Forecasting has been introduced in education settings for a long time; for example, the previous studies have demonstrated the ARIMA can be transferred to tackle specific issues in education [13-15]. The selected ARIMA or ARIMAX with transfer function can deal with universal or multivariable models which are better than their counterparts. In this study, we conduct the cross correlation function (CCF) to determine the relationship between the numbers of students and teachers. If the relationship exists, we determine whether the student-teacher ratio can be used to interpret the trend in future with ARIMA or ARIMAX.

2.1. **The logic of predicted model building.** In this study, we consider the CPSS and CPST for their cross correlation function. CCF is the first step to check if the CPSS and CPST with concurrent relationships for building ARIMAX model. When the CCF does not exist in the two series, we check STR series data. Since the trend of STR is integrated by CPSS and CPST, it may reflect the characteristics of both series data sets. Finally, we can find a fitted model to estimate the trend of student-teacher ratio in the primary education. The framework of model building is presented in Figure 2.

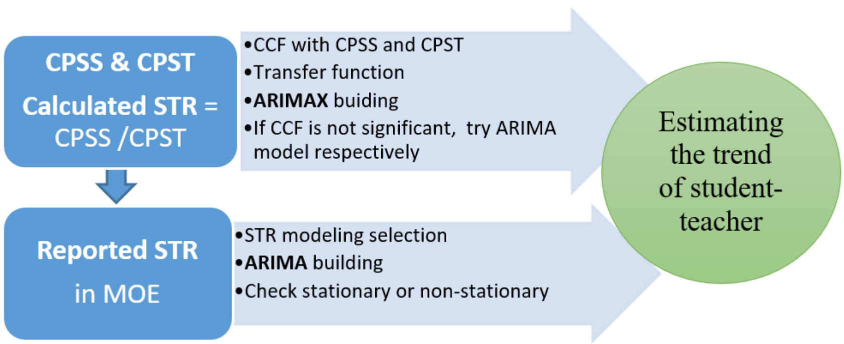


FIGURE 2. The framework of predicting model building

2.2. **Cross correlation function.** CCF is the degree of similarity between two time series in different time or space which the lag can be considered when time is under investigation. In this study, the trends of student numbers and teacher numbers in the primary education setting have fitted the definition of CCF in order to have further verification. Previous studies have provided useful suggestions to conduct CCF. For example, Mardia and Goodall defined separable CCF as $C_{ij}(X_1, X_2) = \rho(X_1, X_2)a_{ij}$, where $A = [a_{ij}]$ is a $p \times p$ positive definite matrix and $\rho(X_1, X_2)$ is a valid correlation function [16]. For further interpretation, given two processes X_{1t} and X_{2t}, (X_{1t}, X_{2t+k}) is the cross correlation between X_{1t} and X_{2t} at lag k, while, $\rho(x_{2t}, x_{1t+k})$ is the cross correlation between X_{2t} and X_{1t} at lag k [17]. In the case of X and Y, the variable X may be cross correlated at different lags of Y, and vice versa. In this study, we propose a way to detect cross correlation coefficients with SPSS (statistic program for social science) program to determine whether the CCF exists in both non-stationary series. Moreover, we can use the following rules to judge the two series which is dependent or independent variable:

- When r_{xy} is positive and significant, x_t is possible as independent variable, while y_t is dependent variable in the model.
- When r_{xy} is significant in lag 0 only, x_t and y_t are concurrently with their impacts. It implies the x_t impacts y_t, while y_t also impacts x_t.
- When r_{xy} is significant with positive and negative values in certain lags, we may assume that x_t impacts y_t, where the impact of y_t will feed back to x_t.

In this case, first we check the series with stationary, then we have the CCF that is the significant cross correlation coefficient having both series data with .05 significant level. When the CCF exists, we conduct the ARIMAX.

2.3. **The process of forecasting.** In this study, the processes of model building are as follows. First, detect if the series of STR is seasonal or non-seasonal. Second, select ARIMA(p, d, q) models by using the differences and visualizations of ACF (autocorrelation function) and PACF (partial autocorrelation function) [18,19]. Third, verify the robustness of the series with the fitted ARIMA model for the next 10 years in terms of the projections from 2019 to 2028. The analysis is carried out using SPSS. Typically, a non-seasonal ARIMA model is classified as an "ARIMA(p, d, q)" model, where: p is the number of autoregressive terms, which represents AR; d is the number of non-seasonal differences needs for stationarity, and two differences are satisfied most of series; q is the number of lagged forecast errors in the prediction equation, which represents MA.

When the difference fits the model building. The fittest model selection will depend on its parameters, BIC (Bayesian information criterion) and Q test. In this study, Box-Pierce Chi-square statistics (Ljung Box test) were used to determine whether the model met the assumptions that the residuals were independent [5]. For Q test, the calculations were listed as follows [10-21]:

$$Q^*(K) = (n-d) \cdot (n-d+2) \cdot \sum_{l=1}^{K} (n-d-l) \cdot r_l^2\,(\hat{a})$$

where n is the sample size, d is the degree of non-seasonal differencing used to transform the series to a stationary one, $r_l^2\,(\hat{a})$ is the sample autocorrelation at lag l for the residuals of the estimated model, and K is the number of lags covering multiple of seasonal cycles, e.g., 12, 24, 36,$\ldots$, for yearly data.

The null hypothesis of the Ljung Box test, H_0, is that our model does not show lack of fit (or in simple terms, the model is just fine). The alternate hypothesis, H_a, is just that the model does show a lack of fit.

3. Results.

3.1. **Concurrent relationship between the trends of students and teachers.** The result of CCF shows the series of primary school students and teachers are only significant in lag 0 (CCF $r = .839$). It implies both series data are not a good fit for transfer function ARIMA model in terms of the fact that ARIMAX did not exist in this case. The result of transformation with natural logarithm and non-seasonal differencing 2 times displays in Table 1 and Figure 3. Table 1 displays the cross correlation coefficients in different lags range from -7 to 7. The result of CCF reveals that the series of primary school students

Table 1. Cross correlations of CPSS and CPST

Lag	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7
Cross correlation	0.357	−0.117	0.171	−0.087	−0.189	−0.141	−0.245	0.839	−0.092	−0.103	−0.311	−0.033	0.113	−0.023	0.337
Std. error	0.128	0.127	0.126	0.125	0.124	0.123	0.122	0.121	0.122	0.123	0.124	0.125	0.126	0.127	0.128

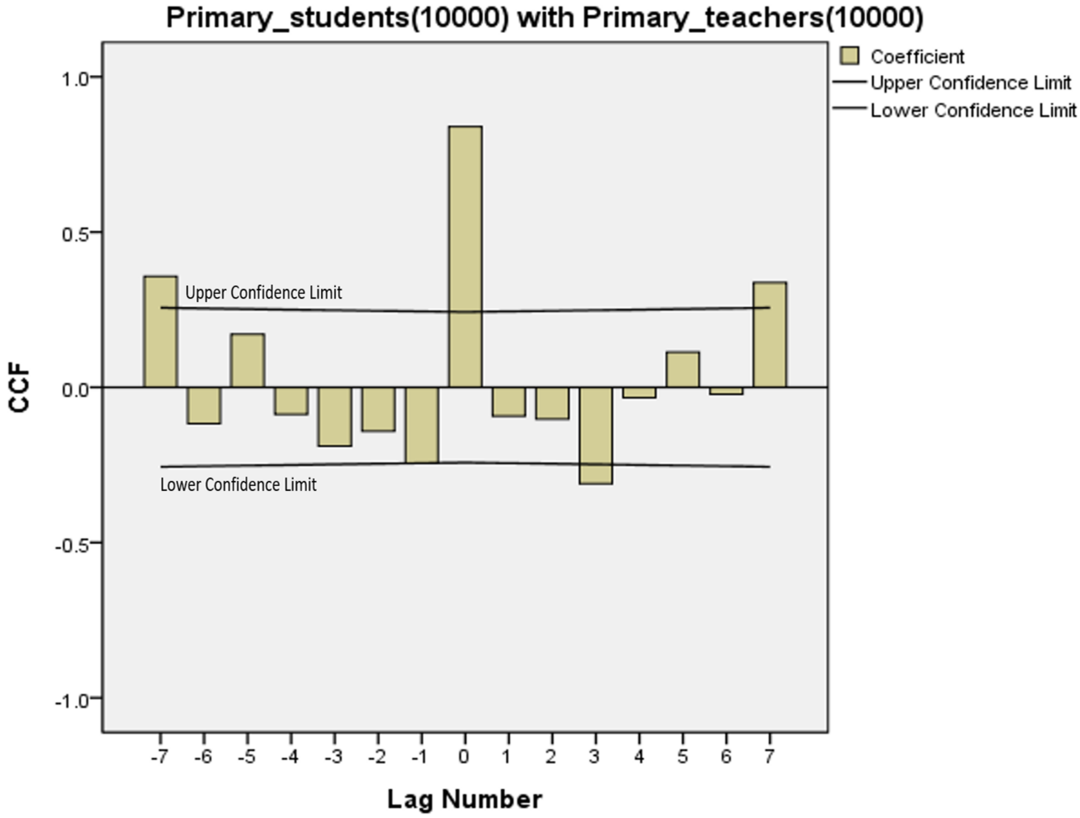


Figure 3. The significance of coefficients in CCF

and teachers did not fit ARIMAX model. The ARIMA model building will conduct for **CPSS** and **CPST** respectively.

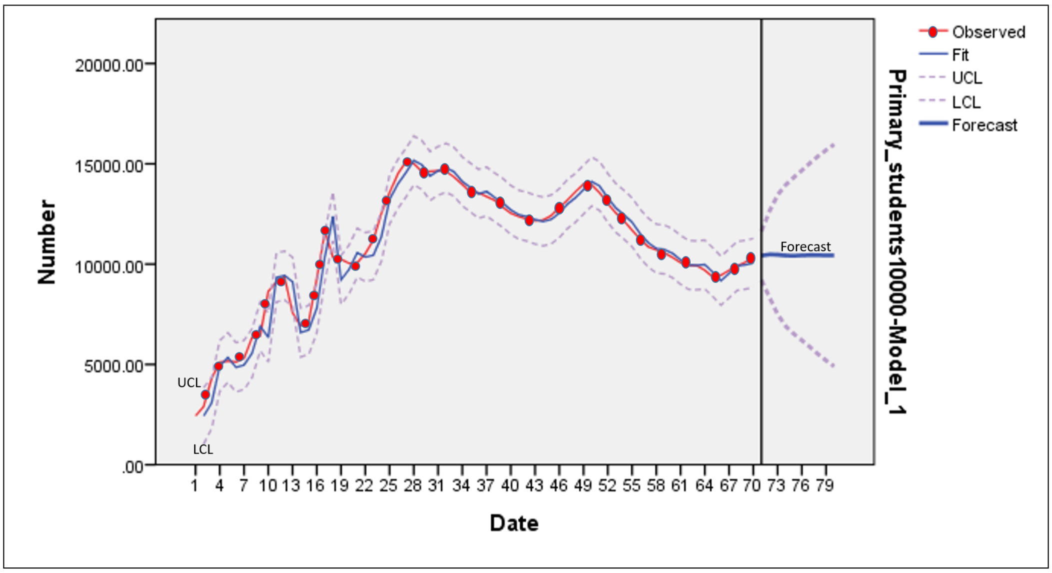
3.2. **ARIMA building for CPSS and CPST.** The SPSS program suggests ARIMA $(2, 1, 2)$ for CPSS and ARIMA$(0, 1, 0)$ for CPST are the fittest models. The predicted values of CPSS and CPST for next decade are listed as Table 2. Based on the predicted values, we found the student numbers will change very slim. While the numbers of teachers will increase significantly in next decade. The plot of CPSS and CPST is displayed in Figure 4.

3.3. **ARIMA building for STR series.** Based on the data of Ministry of Education, the largest number of student in primary schools is 150,941 thousands (1975), and recent lower is 93,605 thousands (2013). Since the student numbers have shown declining in recent years, the numbers of teachers still increase steadily. Will the phenomenon impact the student-teacher ratio in future? The ARIMA$(2, 1, 2)$ suggests trend of CPSS is steady in next decade. The increasing student numbers do not exist significantly in the predicted model. However, the CPST with ARIMA$(0, 1, 0)$ shows it will increase significantly in next decade. The predicted CPSS/CPST will provide a new STR in next decade.

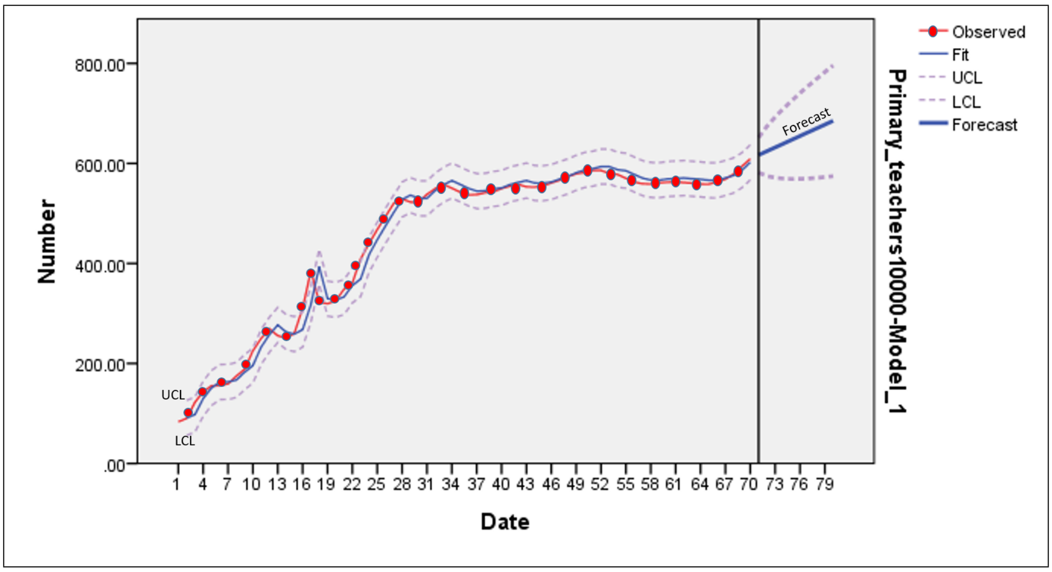
In this case, we also suggest to consider the STR series data directly, and it may become an alternative approach to realize the issue in future. Therefore, we detect the student-teacher ratio series with ARIMA model to build its future trend. ARIMA$(3, 1, 0)$ suggests

TABLE 2. The predicted values with ARIMA(2, 1, 2) for CPSS and ARIMA (0, 1, 0) for CPST

Period	Year	CPSS			CPST		
		Forecast	UCL	LCL	Forecast	UCL	LCL
71	2019	10424.42	11641.68	9207.16	616.81	651.83	581.79
72	2020	10487.93	12636.75	8339.11	624.42	673.95	574.90
73	2021	10471.68	13445.23	7498.13	632.04	692.69	571.39
74	2022	10430.64	13954.81	6906.47	639.66	709.70	569.62
75	2023	10420.59	14318.78	6522.40	647.28	725.58	568.97
76	2024	10437.76	14670.99	6204.54	654.89	740.67	569.12
77	2025	10450.90	15036.51	5865.29	662.51	755.16	569.86
78	2026	10447.77	15382.16	5513.39	670.13	769.17	571.08
79	2027	10439.39	15688.30	5190.48	677.75	782.80	572.69
80	2028	10437.23	15968.11	4906.36	685.36	796.10	574.63



(a)



(b)

FIGURE 4. ARIMA(2, 1, 2) for CPSS (a) and ARIMA(0, 1, 0) for CPST (b)

the RMSE = 1.225, normalized BIC = 0.529 are the smallest. RMSE is measures of error and disregards the complexity of the model. In this case, optimization of RMSE is 1.225 and gives more accurate results, but it could lead to overly complex model that captures too much noise in the data overfitting. In this suggested model, the Ljung-Box Q (18) = 0.068 is also satisfied the white noise test. The limited residuals of ACF and PACF in the model are demonstrated in Figure 5. The testing of parameters for primary student-teacher ratio is displayed in Table 3. The suggested model shows the AR is significant in Lag 1 and Lag 3 respectively with one time difference.

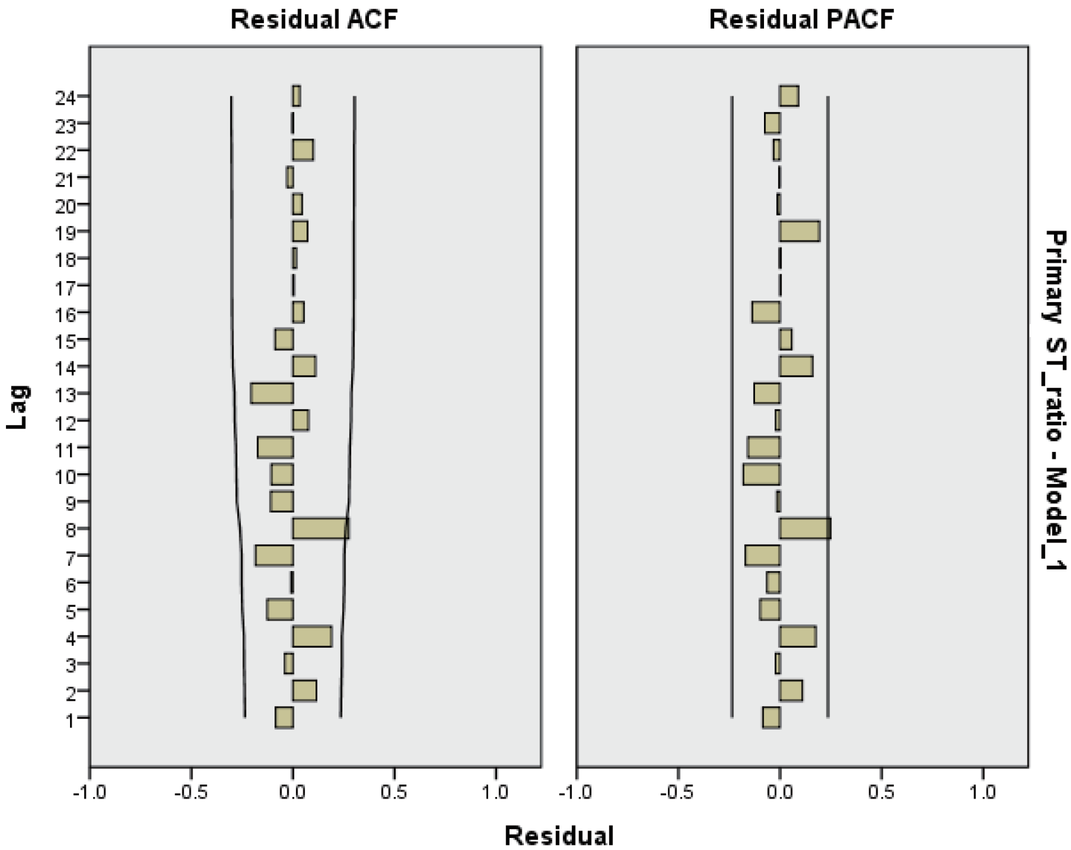


FIGURE 5. The residual of ACF and PACF for building ARIMA$(3, 1, 0)$ for STR

TABLE 3. ARIMA$(3, 1, 0)$ parameters for primary student-teacher ratio

Primary ST_ratio	Model		Estimate	SE	t	Sig.
	AR	Lag 1	0.322	0.112	2.871	0.005
		Lag 3	−0.412	0.107	−3.846	0.000
	Difference		1			

Note. Best-fitting models according to stationary R-squared (larger values indicate better fit).

Comparisons of the predicted STR with ARIMA$(3, 1, 0)$ and STR with predicted CPSS/CPST for next decade are listed as Table 4. The result reveals the STR will increase from 16.95 (in 2019) to 17.13 (in 2028). The plot of ARIMA$(3, 1, 0)$ for building STR with observed and forecast values has been demonstrated in Figure 6. The model suggests the STR in future may experience a little turbulence regardless the students and teachers increasing rapidly. While the CPSS and CPST based on ARIMA$(2, 1, 2)$ and ARIMA$(0, 1, 0)$ suggest the STR will decrease from 16.90 in 2019 to 15.23 in 2028. There is a gap between the two different ways to predict STR. The results provide useful information for related policy making.

TABLE 4. Forecast for STR with different models

Year	STR forecasts with ARIMA(3, 1.0)	CPSS/CPST based on ARIMA(2, 1, 2) and ARIMA(0, 1, 0)
2019	16.95	16.90
2020	17.02	16.80
2021	17.07	16.57
2022	17.1	16.31
2023	17.09	16.10
2024	17.09	15.94
2025	17.08	15.77
2026	17.1	15.59
2027	17.11	15.40
2028	17.13	15.23

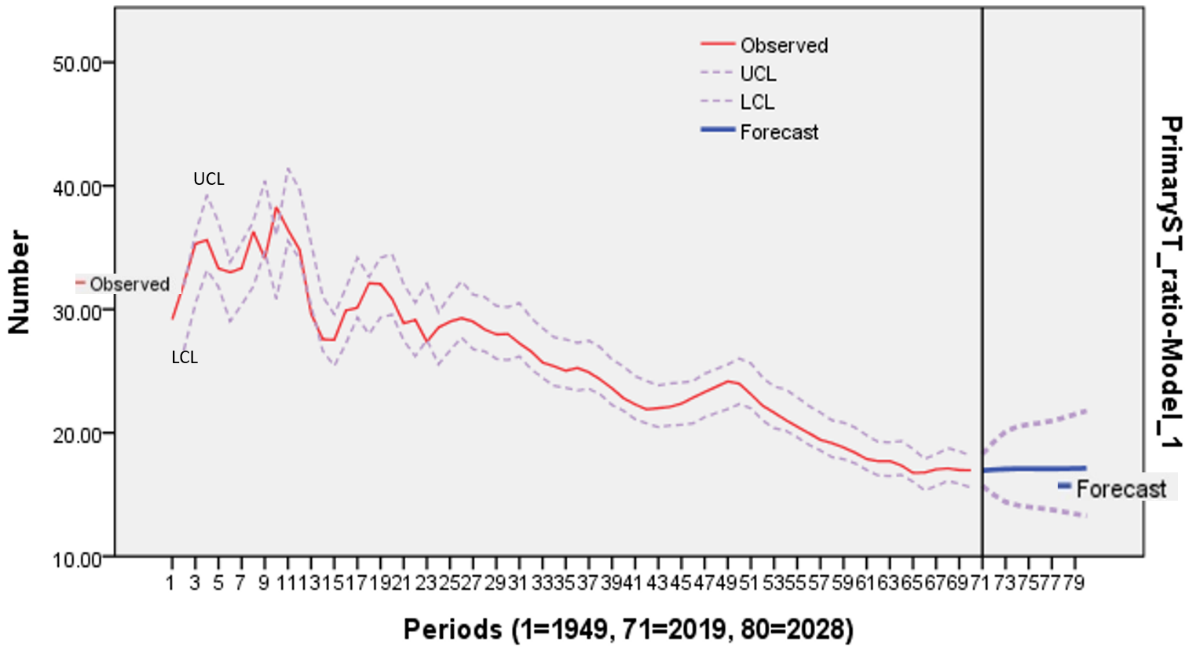


FIGURE 6. The predicted values for STR from 2019 (period 71) to 2028 (period 80)

4. **Conclusions.** Based on the global context and the systematic data analysis, this study tackles the STR issue in a declining student enrollment system. The findings reveal the calculated STR with the trend of fluctuating student and teacher numbers will decline in future. Although the quantitative approach was limited by the data set, the explanation of the forecasting provides a longitudinal perspective for reviewing the trend of STR in primary education. This design model can be used to detect the declining of student numbers and the increasing phenomenon of teacher numbers in primary education. In global context, the ratio of students to teaching staff in primary education is on average 15 in OECD countries and on average 14 in EU22, and China is approaching to that standard as our predication. Considering the characteristics of STR, it is a practical indicator to review the quality of primary education. The result of forecasting provides useful policy intervention information for balancing the numbers of students and teachers. For better enhancing quality of teaching, we suggest: First, more specific to review the differences in cities and rural areas to highlight the STR issue; Second, reducing the STR in future to fit average level in the OECD counties; Third, building a long term review mechanism

to prompt the issue is needed, for example, the mechanism is based on annual data set automatically.

For further studies, we suggest both ARIMA and ARIMAX are useful tool to explore time series data set. The design of this study can be extended to similar settings to tackle the related predicting problems.

REFERENCES

[1] X. Liang, *China: Challenges of Secondary Education. World Bank, Education*, 2001.
[2] National Bureau of Statistics of China, *China Statistical Yearbook*, 2016.
[3] National Bureau of Statistics of China, *Numbers of Students and Teachers in Primary Education*, http://www.stats.gov.cn/tjsj/, 2019.
[4] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications with R Examples*, 4th Edition, Springer, Cham, Switzerland, 2017.
[5] B. Nath, D. S. Dhakre and D. Bhattacharya, Forecasting wheat production in India: An ARIMA modelling approach, *Journal of Pharmacognosy and Phytochemistry*, vol.8, no.1, pp.2158-2165, 2019.
[6] C. Yuan, S. Liu and Z. Fang, Comparison of China's primary energy consumption forecasting by using ARIMA model and GM$(1, 1)$ model, *Energy*, vol.100, pp.384-390, 2016.
[7] S. J. Wu, D. F. Chang and H. Hu, Detecting the issue of higher education over-expanded under declining enrollment times, *Higher Education Policy*, https://doi.org/10.1057/s41307-019-00163-z, 2019.
[8] N. Achille, S. Haberman and G. Consigli, A multivariate approach to project common trends in mortality indices, *SSRN Electronic Journal*, DOI: 10.2139/ssrn.3149989, 2018.
[9] M. B. Chamlin and B. A. Sanders, Social policy and crash fatalities: A multivariate time series analysis, *Journal of Crime and Justice*, vol.41, no.3, pp.322-333, 2018.
[10] D.-F. Chang and K. L. Lai, Trajectory of the population dependency index by using ARIMA models, *ICIC Express Letters, Part B: Applications*, vol.10, no.3, pp.195-202, 2019.
[11] J. Ruan and Z. Lu, A self-adaptive spatial-temporal correlation prediction algorithm to reduce data transmission in wireless sensor networks, *International Journal of Innovative Computing, Information and Control*, vol.14, no.3, pp.997-1013, 2018.
[12] OECD, *OECD Education at a Glance 2017*, OECD Indicators, Paris, France, 2017.
[13] D.-F. Chang, Effects of higher education expansion on gender parity: A 65-year trajectory in Taiwan, *Higher Education*, vol.76, no.3, pp.449-466, 2017.
[14] D.-F. Chang and H. Hu, Mining gender parity patterns in STEM by using ARIMA model, *ICIC Express Letters, Part B: Applications*, vol.10, no.2, pp.105-112, 2019.
[15] H.-C. ChangTzeng, D.-F. Chang and Y.-H. Lo, Detecting concurrent relationships of selected time series data for ARIMAX model, *ICIC Express Letters, Part B: Applications*, vol.10, no.10, pp.937-944, 2019.
[16] K. V. Mardia and C. R. Goodall, Spatial-temporal analysis of multivariate environmental monitoring data, in *Multivariate Environmental Statistics*, G. P. Patil and C. R. Rao (eds.), North-Holland, Amsterdam, The Netherlands, 1993.
[17] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 5th Edition, Wiley, New York, NY, 2015.
[18] R. Davies, T. Coole and D. Osipyw, The application of time series modelling and Monte Carlo simulation: Forecasting volatile inventory requirements, *Applied Mathematics*, vol.5, no.8, pp.1152-1168, DOI: 10.4236/am.2014.58108, 2014.
[19] P. R. Junior, F. L. R. Salomon and E. de O. Pamplona, ARIMA: An applied time series forecasting model for the Bovespa stock index, *Applied Mathematics*, vol.5, no.21, pp.3383-3391, DOI: 10.4236/am.2014.521315, 2014.
[20] G. M. Ljung and G. E. P. Box, On a measure of lack of fit in time series models, *Biometrika*, vol.65, no.2, pp.297-303, 1978.
[21] R. Nau, *ARIMA Models for Time Series Forecasting*, https://people.duke.edu/~rnau/411arim.htm, 2014.