ICIC Express Letters
Part B: Applications
Volume **11**, Number **2**, February **2020**

ICIC International ©2020 ISSN 2185-2766
pp. **121–128**

FORECASTING WITH ARIMAX MODELS FOR PARTICIPATING STEM PROGRAMS

Dian-Fu Chang[1], Chia-Chi Chen[2] and Angel Chang[3,*]

[1]Graduate Institute of Educational Policy and Leadership
[2]Doctoral Program of Educational Leadership and Technology Management
Tamkang University
No. 151, Yingzhuan Road, Tamsui District, New Taipei City 25137, Taiwan
140626@mail.tku.edu.tw; sophiabv03@gmail.com

[3]College of Humanities and the Arts
San Jose State University
1 Washington Sq, San Jose, CA 95192, USA
*Corresponding author: hcachang@ucdavis.edu

Received August 2019; accepted November 2019

Abstract. *Many studies have examined different fields of higher education expansion as well as the understanding of expansion through the relationship between higher education and other academic fields. This study examined how the expansion of higher education impacts STEM (science, technology, engineering and mathematics) programs and differentiates in the trajectories of Taiwan. This study aims to explore the expansion phenomenon related to the enrollment in STEM in expanding higher education. We used the classical ARIMA model to provide forecasts for the Ministry of Education (MOE) dataset. We then implemented ARIMAX (a multivariate autoregressive integrated moving average model) method to deal with the two concurrent series. The data source of this study, the time series data of student enrollment in the STEM programs and total student numbers (1950 to 2018), retrieved from MOE, Taiwan. We conducted the cross-correlation function to check the relationships between the series. We employed the ARIMAX methods to select the best fit model to predict student enrollment in STEM programs. The result revealed the selected ARIMAX(1,2,1) works well to establish the best fit model to predict enrollment in STEM programs. This finding provided implication to educational policy makers to implement the innovative STEM programs.*

Keywords: ARIMA, ARIMAX, Cross-correlation function, Higher education, STEM, Transfer function

1. **Introduction.** Global higher education systems constantly face the balance of quality, equity, and resource [1], and higher education systems in Taiwan are no exception [2]. In the last few decades, higher education systems in Taiwan expanded in terms of the access and geographical locations of the universities. In addition, the higher education systems also expanded access to nontraditional students such as low-income students [3-5].

According to data from the UNESCO (The United Nations Educational, Scientific and Cultural Organization) Institute for Statistics (2018), the gross entrance ratio (GER) in high income countries moved to a universal stage in 1993 [6]. The results reflect outcomes for globalization but also improvement made in several different facets such as increased participation of women in higher education, specifically the enrollment and completion rates. In Taiwan, the GER reached 75% in 2011 while most of the middle-income countries moved to that stage in 2001 [7].

Currently, one-third of the world's college-age population participates in higher education [8]. Thus, higher education in Taiwan has expanded rapidly in the previous three

DOI: 10.24507/icicelb.11.02.121

decades. From 1976 to1999, the enrollment increased from 299,486 to 576,623 students while the GER of higher education system rose from 15 percent to 50 percent within 23 years in [9]. In 1999, the popularization of education had increased student enrollment two times more compared to that in 1976. According to the 2015 Education Statistical Indicators, the tertiary education GER hit 83.88 percent in 2013, which was higher than that of most other Asian countries [10,11]. The absolute limits to expansion have been rarely discussed directly in higher education literature. However, much research has approached higher education expansion indirectly, either by debating whether such limits have already been reached (i.e., over-qualification) or through theorization.

This study may provide an alternative way to tackle the issue. Meanwhile, the birthrate in Taiwan has decreased from 328,461 in 1974 to 196,973 in 2016, showing a 40 percent decrease according to data from the Ministry of Interior [12]. Under the declining trend, previous studies have indicated that many private higher education institutions have found themselves confronted with a serious shortage of student recruitment [13]. Even though the STEM (science, technology, engineering and mathematics) programs still attract a large part of enrolled students, we are concerned that such decline might impact the enrollment in STEM.

In this study, we selected STEM programs in Taiwan as the research target to detect how the time series data work within the higher education system. Through examining the enrollment, resources, and equity of the STEM program, we might be able to understand the current general Taiwanese higher education. Given this purpose, this study explores the following research questions:

a) What kind of relationship is between the series of STEM and expanding higher education system?

b) How the selected model could interpret the two concurrent series?

c) Which predicted model fits to the phenomenon in the future?

To answer the questions, this paper begins with the method section which introduces research framework, definition of target series data, and the statistical process. Then, we present the related findings in the result section. Finally, the conclusion will be drawn.

The application of time series analysis can be applied to the diverse fields; one of the systematic models, autoregressive integrated moving average (ARIMA) model, deals with time-correlated modeling and forecasting [14]. This is useful for analyzing the joint behavior of two stationary series whose behavior may be related in some unspecified way [15,16]. To extend the applications, this study conducted the CCF, transfer function and multivariate autoregressive integrated moving average (ARIMAX) to build the best fit model in order to predict and provide interpretation of enrollment trends in higher education.

2. **Method.**

To tackle the expanding issue, we draw a research framework to rationalize the process of study. First, we collect the data sets to fit the criteria of conducting ARIMAX. Second, the algorithm of statistical process will be addressed.

2.1. **Research framework.**

Figure 1 describes how we conduct the transfer function ARIMAX with the target series data. In the first step, we collected the data from the Ministry of Education and integrated the data sets with a reasonable and meaningful way. In the second step, we check the data sets to fit the CCF requirement, if the series with high CCF, we go through the ARIMAX. The auto-correlation and cross-correlation functions (CCF) served as tools to clarify relationships that may occur within and between time series at various lags. The ARIMAX model can be viewed as a multiple regression model with one or more autoregressive (AR) terms and/or one or more moving average (MA) terms. The proposed predicting model referred to ARIMAX, and the X stands for exogenous [17]. In the third step, we build the transfer function models with difference,

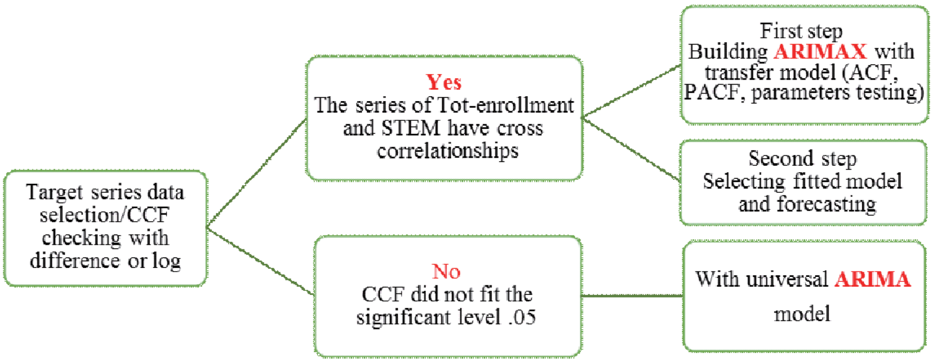
Target series data selection/CCF checking with difference or log
Yes
The series of Tot-enrollment and STEM have cross correlationships
No
CCF did not fit the significant level .05
First step
Building ARIMAX with transfer model (ACF, PACF, parameters testing)
Second step
Selecting fitted model and forecasting
With universal ARIMA model

FIGURE 1. The research framework

log, autocorrelation function (ACF) or partial autocorrelation function (PACF). Finally, based on the estimation of parameters, we select the fittest model for interpreting the series.

2.2. **Target series data.** The date which is used is in the form of yearly 69 period data from 1950 to 2018 [18]. The data sources include online version and document format data in previous publications by Ministry of Education. The STEM programs have shown changing rapidly during these periods. We considered the participation in STEM based on the names of those programs. The data cleaning is a major concern in this study. It takes time to recount the enrollment in STEM during this long journey.

2.3. **Algorithm of statistical process.** To detect the relationship of two series, we conduct CCF to verify the relationships of the series. The CCF has been defined as follows [19,20]:

$$CCF_{XY}(k) = \frac{c_{XY}(k)}{\sqrt{c_{XX}(0)c_{YY}(0)}}$$

where

$$c_{XY}(k) = \begin{cases} \dfrac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(y_{t+k} - \bar{y}), & k = 0, 1, \ldots, n-1 \\ \dfrac{1}{n}\sum_{t=1-k}^{n}(x_t - \bar{x})(y_{t+k} - \bar{y}), & k = -1, -2, \ldots, -(n-1) \end{cases}$$

where $c_{XX}(0)$ and $c_{YY}(0)$ are the sample variances of $\{Xt\}$ and $\{Yt\}$. The CCF calculates the linear correlation between the series, ranging from −1 to 1. In this study, the CCF is conducted by using statistical package for the social sciences (SPSS) package.

To conduct the model building, first, the series will be checked for stationary or nonstationary reasons. Then, we select the fittest model to get estimated parameters. The SPSS program shows both series with log and one difference work well in the CCF. When the CCF has been confirmed, we can assign the dependent (output or responsible) variable and independent variable (input or predictor). Based on the result of CCF, we select STEM students (namely STEM in data file) as dependent variable contemporaneously with total students (namely Tot_enrollment in data file) in the ARIMAX model.

Specifically, this study goes along with ARIMAX as the following steps.

(a) Plot the ACF and PACF of the data and check the series.

(b) Estimate the parameters and test for the significance of the estimates parameters.

(c) Explain why, using the results of parts (a) and (b), it would seem reasonable to differentiate the data prior to the analysis.

(d) Plot the ACF and PACF and check the fitted models. Here, Q-test will be conducted to check whether the null hypothesis of the white noise.

(e) Fit an ARIMAX model.

3. Results.

In this section, we demonstrate the findings of CCF, display the process of ARIMAX model selection, and build the fittest model for predicting the participation in STEM.

3.1. Determination of CCF.

With the log and one difference, we found both series are stationary for conducting CCF. Table 1 displays the cross correlation coefficients in different lags range from -7 to 7. According to 95% significant level, the cross correlation coefficients among lags -7 to 4 are all significant, see Table 1 and Figure 2. It implies both series with high correlation and fit to build predict models with transfer function ARIMAX.

Table 1. Cross-correlation function with total and STEM students (1950-2018)

Lag	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
Cross correlation	0.418	0.456	0.516	0.586	0.651	0.732	0.828	0.903	0.78	0.595	0.44	0.33	0.22	0.119	0.053
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

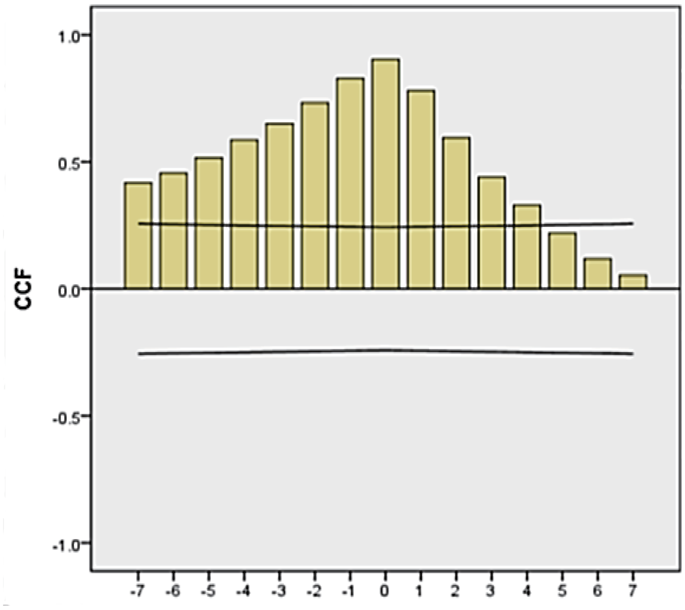

Figure 2. Testing the significance of CCF

3.2. Selection of the fittest ARIMAX.

Based on the previous information, we select the ARIMAX(1,2,1) to predict the relationship with total student enrollment and that of STEM programs. The ACF plot reveals the data are stationary because the autocorrelations are all zero; indicate the random error. Similarly, the PACF plot indicates that no significant spike of the lags, as shown in Figure 3. As such, the model design is a good fit for this study.

Table 2 displays details of the fittest statistics in ARIMAX(1,2,1). For example, the mean of smooth R^2 is .813; the mean of RMSE (root mean square error) is 6214.742; the MAPE (mean absolute percentage error) is 2.156; the mean of BIC (Bayesian information criterion) is 17.919. The percentages from 5 to 95 are also shown stability in this model. The parameters of ARIMAX(1,2,1) demonstrate that "STEM enroll students" with log and one difference is significant in its constant and AR(1) terms. Moreover, the "Total Enrollment" works well with two difference and lag 1 as the denominator in the selected model. While the "Total Enrollment" as the numerator with lag 1 is also suitable in the model. The details of results are presented in Table 3. The result of the visualized plot for observed and fittest predicted values is shown in Figure 4. Based on the graphs and tables, the result reveals the predicted participation in STEM program may decrease in this model.

According to Ljung-Box Chi-square statistics, we found the model meets the assumptions that residual are independent. Technically, a significant level of .05 (denoted as

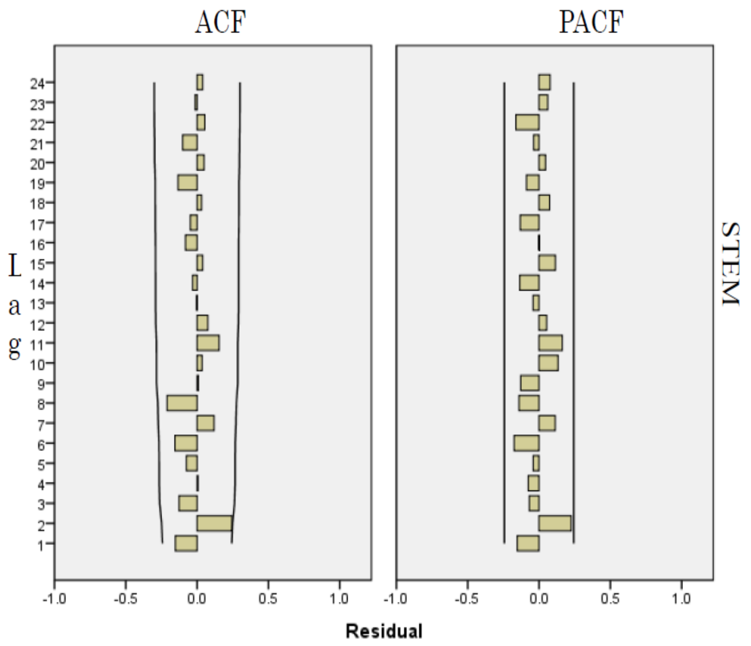

Figure 3. ACF and PACF for transfer function ARIMAX(1,2,1)

Table 2. The fittest statistics of ARIMAX(1,2,1)

Fittest statistics	Mean	Percentage						
		5	10	25	50	75	90	95
Smooth R^2	.813	.813	.813	.813	.813	.813	.813	.813
R^2	.999	.999	.999	.999	.999	.999	.999	.999
RMSE	6214.742	6214.742	6214.742	6214.742	6214.742	6214.742	6214.742	6214.742
MAPE	2.156	2.156	2.156	2.156	2.156	2.156	2.156	2.156
MaxAPE	10.789	10.789	10.789	10.789	10.789	10.789	10.789	10.789
MAE	4315.001	4315.001	4315.001	4315.001	4315.001	4315.001	4315.001	4315.001
MaxAE	15345.629	15345.629	15345.629	15345.629	15345.629	15345.629	15345.629	15345.629
Std. BIC	17.919	17.919	17.919	17.919	17.919	17.919	17.919	17.919

Table 3. The fittest ARIMAX(1,2,1) based on standardized BIC

Model					Estimate	SE	t	p
STEM Model_1	STEM	log	Constant		.071	.058	1.213	.230
			AR	Lag = 1	.955	.054	17.773	.000
			Difference	1				
			MA	Lag = 1	−.431	63.469	−.007	.995
				Lag = 2	.569	36.150	.016	.987
	Total Enrollment	log	Lag	1				
			Numerator	Lag = 0	.113	.066	1.702	.094
				Lag = 1	−.144	.068	−2.120	.038
			Difference	2				
			Denominator	Lag = 1	−.831	.032	−26.294	.000

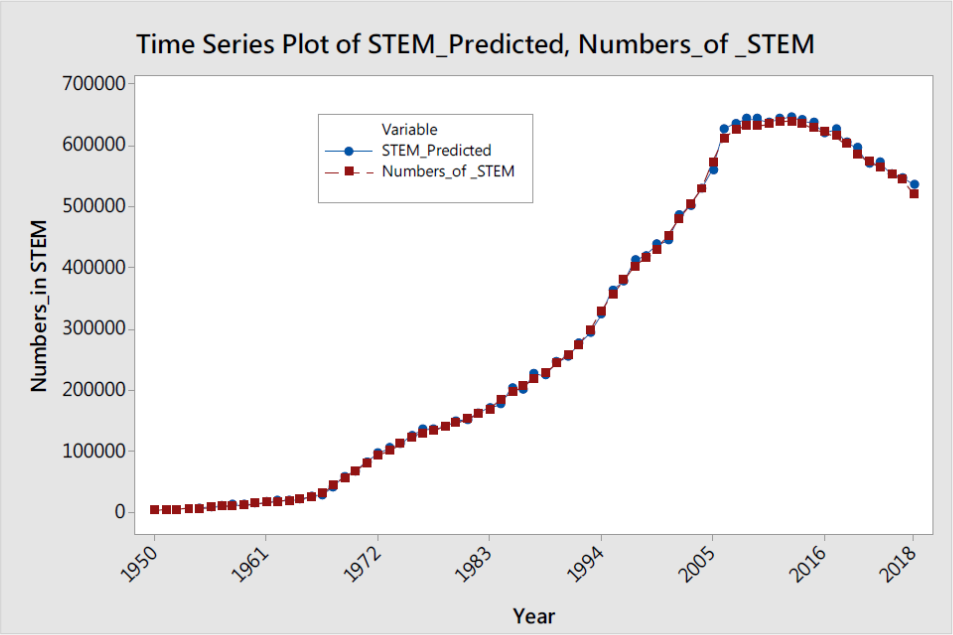

Figure 4. The visualized plot for observed and fittest predicted values

Table 4. Modified Box-Pierce (Ljung-Box) Chi-square statistic

Lag	12	24	36	48
Chi-square	15.81	21.79	29.32	30.70
DF	9	21	33	45
p-value	0.071	0.412	0.651	0.949

α) works well in lag 12, 24, 36 and 48. This statistics of the *p*-value for the Ljung-Box Chi-square statistics are all greater than .05, see Table 4.

3.3. **Forecasting for participation in STEM from period 2019 to 2028.** The forecasting 10 years ahead in series of STEM is demonstrated in Table 5. The results reveal the predication can be used to interpret the STEM's program within the expansion system. The participation in STEM programs is shown accompanied with the expansion in the higher education system. The prediction will follow the participation in higher education; it implies that STEM will decrease steadily in future.

Table 5. Forecasts from period 70-79 (2019-2028) with STEM

Period	Year	Forecast	95% Limits	
			Lower	Upper
70	2019	527414	503515	551313
71	2020	517345	479277	555412
72	2021	507551	447545	567557
73	2022	497307	414920	579693
74	2023	487052	378568	595535
75	2024	476519	340387	612650
76	2025	465871	299642	632099
77	2026	455008	256970	653047
78	2027	443992	212203	675781
79	2028	432785	165593	699977

4. **Conclusion.** This study took Taiwan's higher education participation as a target to tackle the trend of STEM participation and interpret the meanings of findings. This study provided an example of analyzing two time series data sets in a specific higher education system with ARIMAX. The selected ARIMAX model can be used to project the change of STEM enrollment in future. The results reveal the participation of STEM programs may be consistent with higher education expansion. The higher education expanding in the system has shown a new high for couple years ago. In the future, the findings reveal the trend of STEM will decrease steadily. Based on the findings of previous studies, the higher education system has faced the new crisis of declining birthrate. The findings may provide useful information for related policy makers to adjust their enrollment.

To sum up, the ARIMAX works better than that of universal ARIMA model in this case study. The ARIMAX has been used in the frequency domain and other domains. For instance, it can be extended to the classical techniques such as ANOVA and principal component analysis to the multivariate time series cases. In further studies, this study suggests selected fitted series data and using multivariate time series analysis to tackle related issues in other settings.

REFERENCES

[1] S. Ilie and P. Rose, Is equal access to higher education in South Asia and sub-Saharan Africa achievable by 2030?, *Higher Education*, vol.72, no.4, pp.435-455, 2016.

[2] D. F. Chang, F.-Y. Nyeu and H.-C. Chang, Balancing quality and quantity to build research universities in Taiwan, *Higher Education*, vol.70, no.2, doi: 10.1007/s10734-014-9841-y, 2015.

[3] F. M. Msigwa, Widening participation in higher education: A social justice analysis of student loans in Tanzania, *Higher Education*, vol.72, no.4, pp.541-556, 2016.

[4] R. Schendel and T. McCowan, Expanding higher education systems in low- and middle-income countries: The challenges of equity and quality, *Higher Education*, vol.72, no.4, pp.407-411, 2016.

[5] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*, Springer Texts in Statistics, Springer, 2005.

[6] UNESCO Institute for Statistics, *UIS statistics – Gross Enrolment Ratio by Level of Education*, http://data.uis.unesco.org/, 2018.

[7] S. Marginson, The worldwide trend to high participation higher education: Dynamics of social stratification in inclusive systems, *Higher Education*, vol.72, no.4, pp.413-434, 2016.

[8] 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, 2018.

[9] Department of Statistics, Ministry of Education, *Educational Statistics (2014 Edition) Excel File*, http://stats.moe.gov.tw/files/ebook/Education_Statistics/103/103edu_EXCEL.htm, 2016.

[10] Ministry of Interior, *The Main Directory of Dynamic Query Statistics 2018*, http://statis.moi.gov.tw/micst/stmain.jsp?sys=100, 2018.

[11] D. F. Chang and Y. L. Huang, Detecting the effect of policy intervention for oversupply higher education system, *ICIC Express Letters, Part B: Applications*, vol.8, no.11, pp.1489-1495, 2017.

[12] M. A. Trow, *Problems in the Transition from Elite to Mass Higher Education*, Carnegie Commission on Higher Education, Berkeley, CA, 1973.

[13] V. Hohreiter et al., Cross-correlation analysis for temperature measurement, *Meas. Sci. Technol.*, vol.13, pp.1072-1078, 2002.

[14] M. G. Olsen and R. J. Adrian, Out-of-focus effects on particle visibility and correlation in microscopic particle image velocimetry, *Exp. Fluids*, vol.29, pp.166-174, 2000.

[15] L. Dannecker, *Energy Time Series Forecasting*, Springer Vieweg, Wiesbaden, Germany, 2015.

[16] P. Aboagye-Sarfo, Q. Mai, F. M. Sanfilippo, D. B. Preen and L. M. Stewart, A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia, *Journal of Biomedical Informatics*, vol.57, no.10, pp.62-73, 2015.

[17] E. Ekheden and O. Hösjer, Multivariate time series modeling, estimation and prediction of mortalities, *Insurance: Mathematics and Economics*, vol.65, no.11, pp.156-171, 2015.

[18] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications with R Examples*, 4th Edition, Springer, Cham, Switzerland, 2017.

[19] Ministry of Education, *Statistics for Higher Education 2018*, https://stats.moe.gov.tw/bookcase/Higher/108/index.html#p=1, 2018.

[20] C. Chatfield, *The Analysis of Time Series: An Introduction*, Chapman and Hall, London, 1996.

[21] E. M. Souza and V. B. Felix, Wavelet cross-correlation in bivariate time-series analysis, *Trends in Applied and Computational Mathematics*, vol.19, no.3, pp.391-403, 2018.