ATLANTIS PRESS

Advances in Computer Science Research, volume 93

2nd International Conference on Mathematics, Modeling and Simulation Technologies and Applications (MMSTA 2019)

Predicting the Participation in Social Science under Expanding System by Using ARIMAX

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

[1]Graduate Institute of Educational Policy and Leadership, Tamkang University, Taiwan
[2]Doctoral Program of Educational Leadership and Technology Management, Tamkang University, Taiwan
[*]Corresponding author

***Abstract*—This study aims to predict the participation pattern related to the social science programs within a high participated higher education system. Taking the expanding higher education system in Taiwan as an example, we collected the series data with student numbers in social science programs and total student enrollment numbers by using the annual statistics report (1950 to 2017) from Ministry of Education. Considered the concurrent series did not fit the classical ARIMA (autoregressive integrated moving average) model, this study selected transfer function in terms of multivariate autoregressive integrated moving average (ARIMAX) models to deal with the target series. First, we applied the cross correlation function to check the relationships between the series. Second, we select the ARIMAX with transfer function to verify the fittest predicting model. The result reveals the selected ARIMAX(1,1,1) model works well for predicting the trend of social science participation in future. This study provides an example to tackle two series variables in ARIMAX process in higher education settings. The finding suggests useful information for related policy makers to renovating the social science programs.**

Keywords—ARIMA; ARIMAX; cross correlation function; higher education; social science; transfer function

I. INTRODUCTION

Various studies deal with series data from different strategies in time series field. For example, the dynamic series data sets are unique cases, they usually contain several observable variables that exhibit long-range dependence or multifractal nature. Previous studies have demonstrated that turbulent flows, velocities, temperatures, stock markets, and concentration fields are embedded in the similar space as joint multifractal measures [1-5]. Reviewed previous literature, we found the autocorrelation and cross correlation functions (CCF) are useful for analyzing the joint behaviors of two stationary series whose behaviors may be related in some unspecified ways [6]. To extend the applications, this study conducted the CCF, transfer function and multivariate autoregressive integrated moving average (ARIMAX) to select the fittest predicting model for interpreting the specific participation in higher education system. These series may accompany with various observable variables. In this sense, the proposed predicting model referred to as an ARIMAX, the X stands for exogenous term [6].

Higher education systems have formatted a rational process of expansion in the last decades. Like Marginson's argument, the expansion process has extended to most middle-income countries and to a significant number of low-income ones [7-9]. According to data from the UNESCO Institute for Statistics (2018), the gross entrance ratio (GER) in high income countries has moved to a universal stage in 1993. Some countries reached 75% in 2011, while the GER in most of the middle-income countries moved to the mass stage in 2001 [10]. Currently, about one third of the world's college age population participates in higher education [11]. Higher education in Taiwan has expanded dramatically in the previous decades. An overview of higher education in the past few decades revealed that the number of students increased from 299,486 (1976) to 576,623 (1999) and the GER (gross enrollment ratio) rose from 15 per cent to 50 per cent within 23 years in the system [12]. The popularization of education has led to a rapid increase in student enrollment, although the figure has leveled off in the last decade. According to the 2015 Education Statistical Indicators, the tertiary education GER hit 83.88 per cent in 2013, which was higher than that in most other Asian countries [12,13]. This is a case of expanding higher education system. Similar phenomenon has shown in other high participation higher education. While much research has approached the topic indirectly, either by debating whether such expansion has already been reached (i.e. over-qualification). We assume the process of expansion, with dynamic factors within the system, may provide meaningful message for studies or related policy makers. The approach of the study may provide an alternative way to tackle the issues. Meanwhile, newborn babies in Taiwan have decreased from 328,461 in 1974 to 196,973 in 2016, showing a 40 per cent drop according to data from the Ministry of Interior [14]. Under the declining trend, previous studies have pointed out that many private higher education institutions have found themselves confronted with a serious shortage of student recruitment [15,16]. According to the above discussions, the system expanding phenomenon might confront a serious crisis related shortage of potential enrollment in future. Under this trend, we found social science has confronted more disadvantaged situation than before. Can social science programs coexisted in the declining trend still survive in future? We concern the systematic expanding or declining might impact the participation in social science programs in the higher education system.

Specifically, we selected social science programs in Taiwan as the research target to detect how the series data work within

the expanding system. Given this purpose, this study explores the following research questions:

a. What kind of relationship between the series of social science program and system expansion in the higher education?

b. Can the ARIMAX be used to interpret the two concurrent series properly?

c. Which predicted model can be used to interpret the phenomenon in future?

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

II. 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. Finally, the fitted model will be selected based on time series' statistical criteria.

A. Research Framework

Figure 1 displays how the transfer function ARIMAX with the target series data will be conducted. First, we collected the data from the Ministry of Education and integrated the data sets with a reasonable and meaningful way. Second, we check the data sets to fit the CCF requirement, if the series with high correlation of CCF, we go through the ARIMAX, otherwise we go through ARIMA model. Third, we may build the transfer function models with difference, 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 data.

B. Target Series Data

We selected the data sets from 1950 to 2017 which cover 68 periods [17]. The data sources cover online version and document format data in previous publications by Ministry of Education. The social science programs have shown changing rapidly during these periods. We considered the participation in social science programs based on college and universities students which including the master and doctoral level of students. The data cleaning work is very important in this study. It takes time to recount the enrollment in social science during this long journey. According to the data, currently social science enrollment in undergraduate level is 38.9%, master level is 34.2%, and doctor level is 18.8% compared with its counterparts (humanity and STEM).

C. Algorithm of statistical process

To detect the relationship of the two series, we conduct CCF to verify the relationships of the series. The CCF has been defined as follows [18,19]:

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

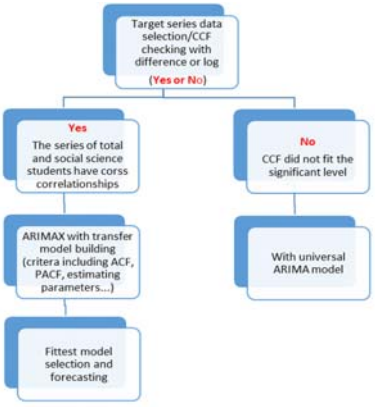

FIGURE I. A FRAMEWORK OF RESEARCH

where

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

$c_{XX}(0)$ and $c_{XY}(0)$ 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 SPSS package, the SPSS programs for CCF are listed as follows:

```
GET DATA /TYPE=XLSX
  /FILE='G:\2019 working papers\For CCF 2019.4.xlsx'
  /SHEET=name '1950-2017'
  /CELLRANGE=full
  /READNAMES=on
  /ASSUMEDSTRWIDTH=32767.
EXECUTE.
DATASET NAME CCF1 WINDOW=FRONT.
SAVE OUTFILE='G:\For CCF\CCF for social_science.sav'
  /COMPRESSED.
CCF
  /VARIABLES=Total Sstotal
  /NOLOG
  /DIFF=1
  /MXCROSS 7.
```

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 will work well in the CCF. When the relationship with 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 social science students (namely Sstolal in data file) as dependent variable contemporaneously with total students (namely Total in data file) in the ARIMAX model. The designed SPSS program for transfer function ARIMAX has listed as follows:

```
PREDICT THRU 77.
TSMODEL
  /MODELSUMMARY  PRINT=[MODELFIT RESIDACF
RESIDPACF] PLOT=[ NORMBIC RESIDACF RESIDPACF]
  /MODELSTATISTICS  DISPLAY=YES MODELFIT=[ SRSQUARE
RSQUARE RMSE MAPE MAE MAXAPE MAXAE NORMBIC]
  /MODELDETAILS  PRINT=[ PARAMETERS RESIDACF
RESIDPACF FORECASTS]  PLOT=[ RESIDACF RESIDPACF]
  /SERIESPLOT OBSERVED FORECAST FIT FORECASTCI FITCI
  /OUTPUTFILTER DISPLAY= [ BESTFIT(PCT=30)]
MODELFIT=NORMBIC
  /SAVE  PREDICTED LCL(LCL) UCL(UCL)
NRESIDUAL(NResidual)
  /AUXILIARY  CILEVEL=95 MAXACFLAGS=24
  /MISSING USERMISSING=EXCLUDE
  /MODEL DEPENDENT=Sstotal INDEPENDENT=Total
    PREFIX='TF_model'
  /ARIMA AR=[1]  DIFF=1  MA=[1]
    TRANSFORM=LN  CONSTANT=YES
  /TRANSFERFUNCTION VARIABLES=Total  NUM=[1]
DENOM=[1] DIFF=1  DELAY=1   TRANSFORM=NONE
  /AUTOOUTLIER DETECT=OFF.
```

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

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

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

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

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

(e) Select the fittest ARIMAX model.

III. RESULTS

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

A. Determination of CCF

With the log and one difference, we found the two series are tended to stationary, see Figure 1. 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 -3 to 5 are all significant, see Figure 3. It implies both series with high correlation and fit to build predict models with transfer function ARIMAX.

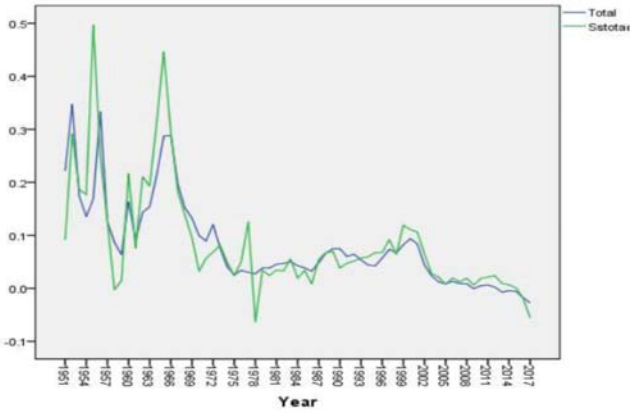

FIGURE II. THE RESULT OF TWO SERIES WITH LOG AND ONE ORDER DIFFERENCE

TABLE I. CROSS CORRELATION FUNCTION WITH TOTAL AND SOCIAL SCIENCE STUDENTS (1950-2017)

Lag	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
Cross correlation	0.04	0.095	0.152	0.261	0.365	0.525	0.705	0.899	0.802	0.659	0.542	0.448	0.343	0.267	0.263
Standardized error	0.129	0.128	0.127	0.126	0.125	0.124	0.123	0.122	0.123	0.124	0.125	0.126	0.127	0.128	0.129

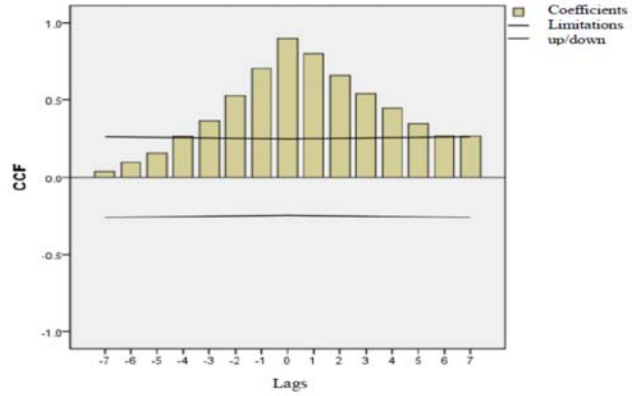

FIGURE III. TESTING THE SIGNIFICANCE OF CCF

B. Selection of the Fittest ARIMAX

Based on previous information, we test the ARIMAX(1,1,1). The ACF plot reveals the data are stationary due to the autocorrelations are all zero or indicative of random error. Similarly, the PACF plot indicates there is no significant spike at the lags, see Figure 4. Both signals are fit the model building.

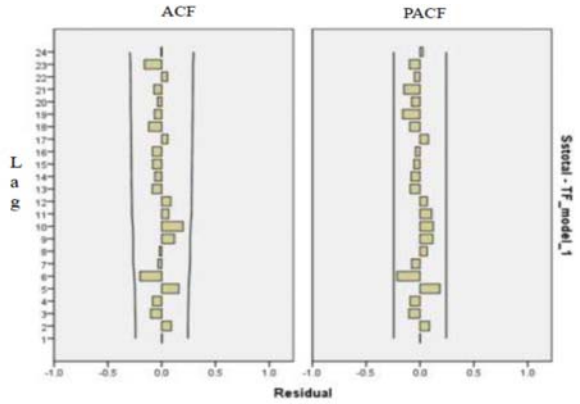

FIGURE IV. ACF AND PACF FOR TRANSFER FUNCTION ARIMAX (1,1,1)

Table 2 displays the details of the fittest statistics in ARIMAX(1,1,1). The results reveal the mean of smooth R^2 is .512, the mean of RMSE (root mean square error) is 6785.518, the MAPE (mean absolute percentage error) is 4.258 and the mean of BIC (Bayesian information criterion) is 18.030. The percentages from 5 to 95 are also shown stability in this model. The parameters of ARIMAX(1,1,1) demonstrate that "Social science students" with log and one difference is significant in its constant and AR(1) terms. Moreover, the "Total students" works well with one difference and lag 1 as the denominator in the selected model. While the "Total students" as the numerator is not suitable in this model based on the result of parameter estimation. The details of results are presented in Table 3. The result of the visualized plot for observed and fittest predicted values shows in Figure 5. In the future, the enrollment of social science programs may decrease as our predicted model.

TABLE II. THE FITTEST STATISTICS OF ARIMAX(1,1,1)

Fittest statistics	Mean	Percentage						
		5	10	25	50	75	90	95
Smooth R^2	.512	.512	.512	.512	.512	.512	.512	.512
R^2	.999	.999	.999	.999	.999	.999	.999	.999
RMSE	6785.518	6785.518	6785.518	6785.518	6785.518	6785.518	6785.518	6785.518
MAPE	4.258	4.258	4.258	4.258	4.258	4.258	4.258	4.258
MaxAPE	28.196	28.196	28.196	28.196	28.196	28.196	28.196	28.196
MAE	4284.890	4284.890	4284.890	4284.890	4284.890	4284.890	4284.890	4284.890
MaxAE	24725.454	24725.454	24725.454	24725.454	24725.454	24725.454	24725.454	24725.454
Std.BIC	18.030	18.030	18.030	18.030	18.030	18.030	18.030	18.030

TABLE III. ESTIMATING THE FITTEST ARIMAX(1,1,1) BASED ON STANDARDIZED BIC

Model					Estimated	SE	*t*	*p*
Social science students TF_model	Social science students	log	Constant		.153	.038	4.048	.000
			AR	Lag=1	.589	.191	3.083	.003
			Difference		1			
			MA	Lag=1	.051	.236	.217	.892
	Total students		Lag		1			
			Numerator	Lag=0	4.812E-007	9.407E-007	2.512	.611
				Lag=1	6.811E-007	1.036E-006	.658	.513
			Difference		1			
			Denominator	Lag=1	.970	.065	14.97	.000

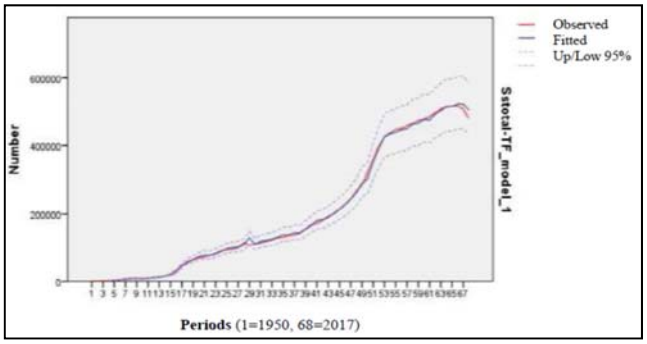

FIGURE V. THE VISUALIZED PLOT FOR OBSERVED AND FITTEST PREDICTED VALUES

IV. CONCLUSION

Taking Taiwan's higher education participation as an example, this study applied data collected from the Ministry of Education to analyze the trend of social science programs and interpret the findings for future development. Long term participation in higher education is difficult to predict without series data and the fittest models. As previous studies, ARIMA models with a statistical package can comfortably transform series data into graphs to detect and create forecasting models, while ARIMAX may provide more accurate estimation with concurrent series. This study provided an example of ARIMAX to analyzing two series data of specific higher education system. The selected ARIMAX(1,1,1) model can be used to project the social science's participation in future. The results reveal the participation of social science programs may consist with higher education expansion. In the future, the finding reveals the trend of social science programs will decrease steadily in future. Since the higher education expanding has shown a new high in the system for couple years ago. We found the system has faced the new crisis of declining birthrate, the findings may provide useful information for related policy makers to adjust related enrollment policy in the system.

Basically, the ARIMAX works well than that of universal ARIMA model in this case study. The ARIMAX can apply in the frequency domain and other domains. For further studies, this study suggests selected fitted concurrent series data and using ARIMAX to tackle the issues in other similar settings.

REFERENCES

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

[2] 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.

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

[4] 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.

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

[6] R. H. Shumway and D. S. Stoffer, Time Series Analysis and Its Applications with R Examples (4th ed.), Cham, Switzerland, Springer, 2017.

[7] 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.

[8] 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.

[9] 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.

[10] UNESCO Institute for Statistics, "UIS statistics-Gross enrolment ratio by level of education", 2018, http://data.uis.unesco.org/

[11] 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.

[12] 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.

[13] Department of Statistics, Ministry of Education, "Educational statistics (2014 edition) excel file, 2016", 2017, http://stats.moe.gov.tw/files/ebook/Education_Statistics/103/103edu_EXCEL.htm

[14] Ministry of Interior, "The main directory of dynamic query statistics 2018", 2018, http://statis.moi.gov.tw/micst/stmain.jsp?sys=100

[15] 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.

[16] 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.

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

[18] C. Chatfield, The Analysis of Time Series: an Introduction, London, Chapman and Hall, 1996.

[19] E. M. Souza and V. B. Felix, "Wavelet cross-correlation in bivariate time-series analysis, Tendencias em Matematica Aplicada e Computacional", vol.19, no.3, pp. 391-403, 2018