ICIC Express Letters
Part B: Applications
Volume **10**, Number **10**, October **2019**

ICIC International ©2019 ISSN 2185-2766
pp. 937–944

DETECTING CONCURRENT RELATIONSHIPS OF SELECTED TIME SERIES DATA FOR ARIMAX MODEL

Hsiao-Chi ChangTzeng[1], Dian-Fu Chang[2] and Yu-Hsin Lo[3,*]

[1]Extension of Education
University of California
1156, High Street, Santa Cruz 95064, USA
hcachang@ucdavis.edu

[2]Graduate Institute of Educational Policy and Leadership
[3]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; *Corresponding author: 807760078@s07.tku.edu.tw

Received May 2019; accepted August 2019

Abstract. *Traditional ARIMA (autoregressive integrated moving average) model works with one series data set, while various cases, with multivariable series data sets, did not fit ARIMA models. Do they have any convenient way to deal with suck kinds of issues? This study provides a practical way to detect two series data sets for ARIMAX (multivariable autoregressive integrated moving average) model building. In our concept framework, three basic components for ARIMAX are proposed including concurrent relationships with series data, verification of cross correlation function, and fitted model for transfer function ARIMAX. This study provides a case study to demonstrate how this detecting process works well for selecting ARIMAX model. The result presents a useful way to interpret the related information from the fittest ARIMAX model.*

Keywords: ARIMA, ARIMAX, Cross correlation function, Time series, Transfer function

1. **Introduction.** ARIMA, standing for autoregressive integrated moving average, is the most popular model for forecasting in time series domains. Previous studies have provided numerous example with this model [1-4]. ARIMA can be made to be "stationary" by differencing (if necessary), perhaps it could be along with nonlinear transformations such as logging or deflating (if necessary) [5]. A random variable, that is a time series, is stationary if its statistical properties are all constant over time. Random-walk and random-trend models, autoregressive models, and exponential smoothing models are all special cases of ARIMA models [5,6]. The world is more complicated than we experienced. The series data are not only random-walk and random-trend with themselves, and there are various concurrent relationships among them. In this sense, it is also called the direct and indirect approaches of forecasting [7]. When we deal with the multivariable series data sets, the traditional ARIMA model cannot fit such kind of cases. Even though the traditional approaches are different, the logic of problem solving for multivariable series may be similar or dissimilar. Moreover, in the social science or humanity field, various cases in time series models are contented with concurrent relationships. When should we select ARIMA or when should we select ARIMAX (multivariable autoregressive integrated moving average)? It is a little confused when we confront such a research topic. This is the reason why we selected this topic for further studies. In this study, we tried to tackle the predicting issues for two series data sets with concurrent relationships. Based on previous studies, the ARIMAX has been proposed as the research target [8]. ARIMAX

DOI: 10.24507/icicelb.10.10.937

model can take the impact of covariates on the forecasting into account, improving the comprehensiveness and accuracy of the prediction [9]. Conducting ARIMAX for predicting, it is a complicated process. Do they have practical ways to deal with such kind of issue? To answer the question, we selected couple series data sets to practice and verify how this process can be conducted. Based on this concern, we try to connect the related detecting process with a reasonable way for ARIMAX model building. Our example will focus on gross domestic products (GDP), population, and enrollment in higher education in the case country. Given this purpose, this study tackles the following research questions.

a) How to select series data sets for building a proposed predicting model?

b) How to detect series data sets with their concurrent relationships for ARIMAX?

c) Which way is more practical for ARIMAX building?

In this study, we begin with the method section, which will address how the related detecting methods can be used. Then, we will address the results with three main topics including defining series data sets, checking cross correlation function (CCF) and building ARIMAX model. Finally, the conclusion will be drawn and the related suggestions will be addressed.

2. **Method.**

This section focuses on how the fitted model can be selected following the logical process. We displayed the theoretical framework to tackle the ARIMAX to proceed the series data. Basically, selecting reasonable data sets is the first important step, then checking CCF, and finally verifying the fitted model.

2.1. **Logics of fitted model.**

The time series data for ARIMAX needed to fit the requirement that both two series data sets are with concurrent relationships. Previous studies suggested the model building at least covering 50 periods [10-12]. This requirement may not be necessary to be an essential in any model. It depends on the property of series data. Moreover, we should consider the meaning of two series data sets for building predicted model. For example, whether X is driven by Y or Y is driven by X. What do they reflect the real situation? The assumption of series relationships needed to be tested with CCF. When both series CCF exist, it can go through transfer function. In the logical framework, the ARIMAX model building is based on the premises. The logical of fittest model selection displays as Figure 1.

2.2. **Definition of time series data.**

Selecting a meaningful time series data is a crucial task for working valuable predictions. In this sense, we consider the data sets should fit the basic requirement for running time series models. For example, the useful and valuable

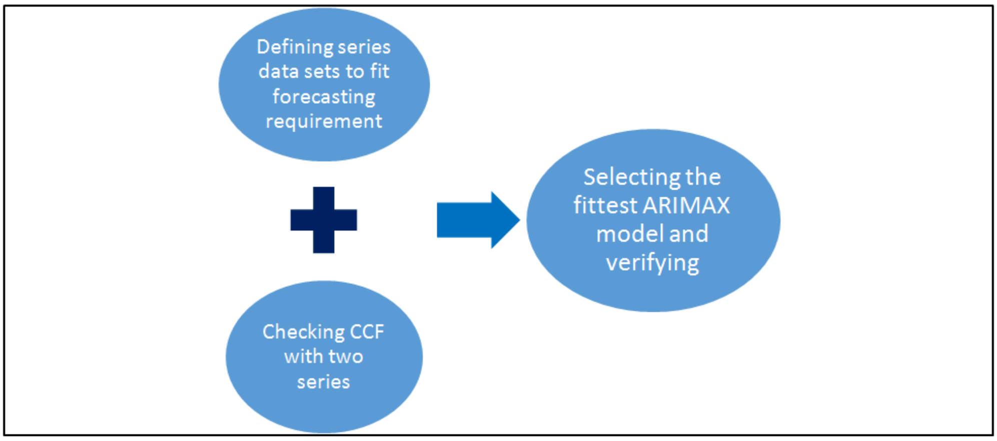
Defining series data sets to fit forecasting requirement
Checking CCF with two series
Selecting the fittest ARIMAX model and verifying

Figure 1. The logical of fittest model selection

duration of data sets is important. Season or non-season series is another factor that should be taken into account in the model building. In this study, we selected GDP, population and enrollment in Taiwan as the target case. The GDP and population series data were selected from National Statistics, Taiwan [13]. The enrollment data sets were derived from Ministry of Education [14]. All the data sets are covering 68 periods (from 1951 to 2018).

2.3. **Cross correlation function.** In an autocorrelation model, it is the cross correlation of a time series while investigating the persistence between lagged times of the same time series or signal. While the CCF is the degree of similarity between two times series in different times or space which the lag can be considered when time is under investigation. The difference between these two times series in different situation like distance, angle, and direction which can be considered while the space is under investigation respectively. To simplify, when we conducted CCF, some basic properties should be considered. As Usoro argued, for X_t and Y_t, the following properties hold in CCF [15]:

a) $\rho_{xy(h)} \leq 1$

b) $\rho_{xy(h)} = \rho_{x(-h)}$

c) $\rho_{xy(0)} \neq 1$

d) $\rho_{xy(h)} = \gamma_{xy(h)} / \sqrt{\gamma_{x(0)}\gamma_{y(0)}}$

Furthermore, Mardia and Goodall defined separable cross correlation function 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]. 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 [11]. 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 proposed a way to detect cross correlation coefficients with their figures to justify whether the CCF exists in both non-stationary series. We can use the following rules to judge the two series which one 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 concurrent 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, the significant cross correlation coefficients were judged by .05 significant level.

2.4. **Building fitted ARIMAX model.** We assume two time series denoted y_t and x_t, which are both stationary. Then, the transfer function model can be written as follows:

$$y_t = C + \nu(B)x_t + N_t$$

where y_t is the output series (dependent variable), x_t is the input series (independent variable), C is constant term, N_t is the disturbance, i.e., the noise series of the system that is independent of the input series. $\nu(B)x_t$ is the transfer function (or impulse response function), which allows x to influence y via a distributed lag. B is a backshift operator and thus we can write as [17-20]

$$\nu(B)x_t = \left(\nu_0 + \nu_1 B + \nu_2 B^2 + \cdots\right) x_t$$

The ARMAX model is quite different from ARIMA model, because it works with two different series x_t and y_t – output series y_t is related to input series x_t. Coefficients ν_j are called impulse response weights, which could be positive or negative. To simplify, the larger the absolute value of any weight ν_j is, the larger is the response of y_t to a change in x_{t-j}. While the output series might not react immediately to a change in input series, thus some initial ν weights may be equal to zero. The number of ν weights equal to zero is called dead time and is denoted as b [17].

3. Results.

The result demonstrates how the CCF works with GDP, enrollment and population. The transfer function was conducted for the series when the CCF is significant. The selected target series work with ARIMAX will be addressed.

3.1. Concurrent relationships verified by CCF.

The finding reveals there is no significant difference between population and GDP. Similarly, there is no significant difference between enrollment and GDP based on their slim coefficients. The different CCFs of the series are displayed in Table 1 and Figure 2. Typically, the lags from 7 to −7 are defaulted in SPSS (statistical programs for social science). In this case, there is no necessary conducting transfer function models.

TABLE 1. Cross correlation coefficients for population, enrollment and GDP

Lag	Population and GDP	Std. error	Enrollment and GDP	Std. error
−7	.027	.129	−.152	.129
−6	.044	.128	−.096	.128
−5	.062	.127	.001	.127
−4	.070	.126	.035	.126
−3	.095	.125	.000	.125
−2	.129	.124	.084	.124
−1	.116	.123	.071	.123
0	.121	.122	.045	.122
1	.085	.123	−.024	.123
2	.152	.124	.105	.124
3	.190	.125	−.012	.125
4	.199	.126	−.018	.126
5	.268	.127	.167	.127
6	.253	.128	.167	.128
7	.276	.129	.129	.129

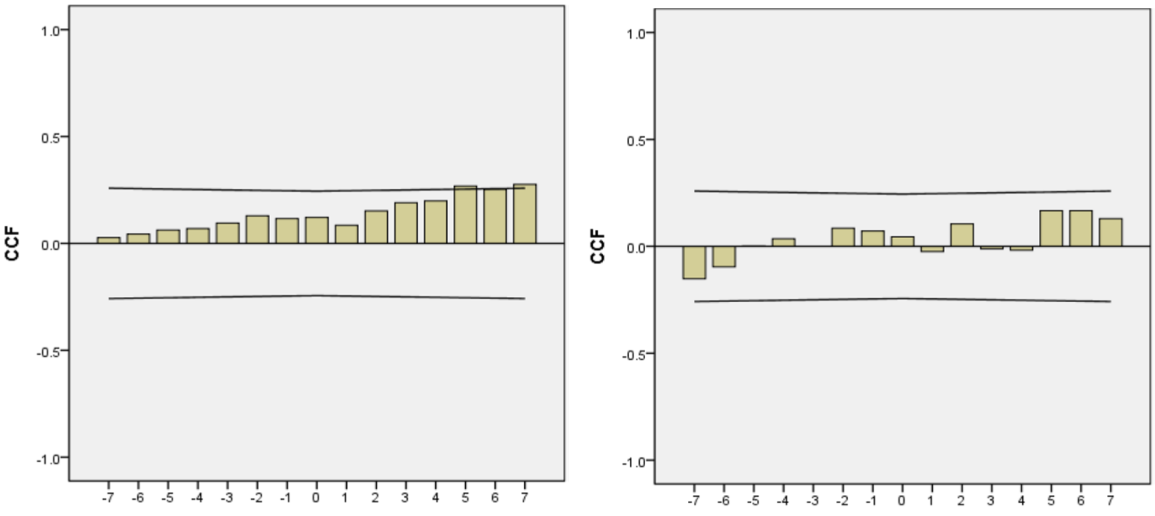

FIGURE 2. Testing the significances of cross correlation coefficients with GDP, enrollment and population (left side is population and GDP, right side is enrollment and GDP)

In contrast, the series between population and enrollment have shown with concurrent relationships. Table 2 and Figure 3 demonstrate there are strong cross relationships between the two series. The coefficients of cross correlation are significant differences from lag 7 to lag -7. Both series data sets fit to ARIMAX model.

Table 2. Coefficients of cross correlation for population and enrollment

Lag	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
Cross correlation	0.515	0.547	0.606	0.67	0.689	0.687	0.708	0.756	0.682	0.66	0.637	0.6	0.513	0.497	0.481
Std. 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

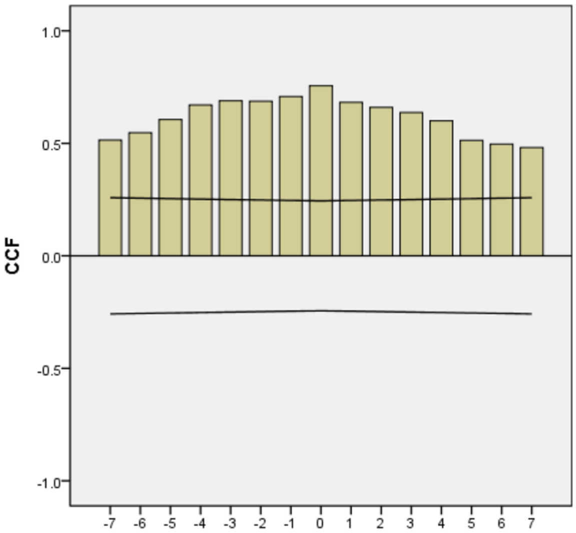

Figure 3. Significant CCF between population and enrollment

3.2. **Selected ARIMAX model building.** We selected ARIMAX$(2, 1, 1)$ with population for forecasting the enrollment. Whether the ARIMAX$(2, 1, 1)$ model fits? The result reveals the smooth R^2 is .733. The standardized BIC is 18.896. The related estimations show ARIMAX$(2, 1, 1)$ with small error in RMSE, MAPE, MAE, and MaxAE. The Ljung-Box Q (18) with df 15 is 28.033 ($p = .021$). It could be a fitted model. The details have been presented in Table 3.

The parameters of ARIMAX$(2, 1, 1)$ demonstrate that the enrollment with log and one difference is significant in its AR(1) terms lag = 1 and lag = 2. Since the CCF has

Table 3. The statistical estimations for ARIMAX$(2, 1, 1)$ model

Estimation for fitted model	ARIMAX$(2, 1, 1)$
Smooth R^2	.733
R^2	1.000
RMSE	9747.535
MAPE	2.164
MaxAPE	18.047
MAE	6606.628
MaxAE	27373.261
Std. BIC	18.896
Ljung-Box Q (18) with df 15	28.033 ($p = .021$)

Table 4. The parameters of ARIMAX$(2, 1, 1)$ based on standardized BIC

Model					Estimate	SE	t	p
Enrollment	Enrollment	log	Constant		−.019	.021	−.902	.371
			AR	Lag = 1	1.313	.237	5.539	.000
				Lag = 2	−.591	.181	−3.275	.002
			Difference	1				
			MA	Lag = 1	.267	.288	.929	.357
	Population	log	Delay = 2					
			Numerator	Lag = 0	3.719	1.538	2.419	.019
				Lag = 1	−2.975	1.418	−2.099	.040
				Lag = 2	−4.130	1.547	−2.671	.010
			Difference	1				
			Denominator	Lag = 1	−.952	.050	−19.152	.000

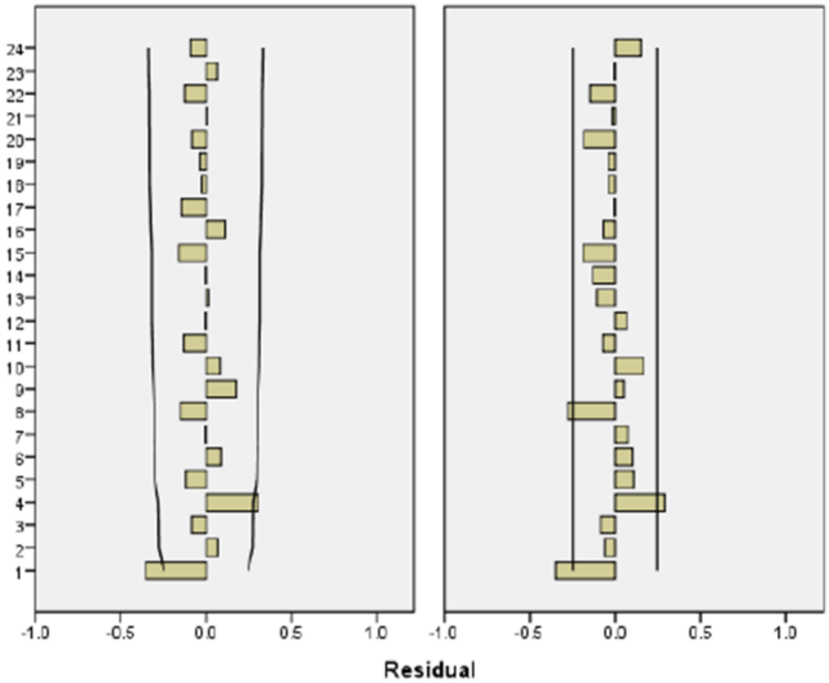

Figure 4. ACF and PACF for ARIMAX$(2, 1, 1)$ model

been fairly well-distributed in this model, the population could be as a numerator or a denominator. The result reveals that the population as a numerator in the model will work in lag 0, lag 1 and lag 2. Moreover, the population works well with one difference and lag 1 as the denominator in the selected model. The details of results are presented in Table 4.

The residual test with ACF (left) and PACF (right) is demonstrated in Figure 4. Both of them fit the assumptions that the white noise is acceptable for the model building.

3.3. **Forecasting enrollment with population.**

Finally, we conducted ARIMAX(2, 1, 1) model to forecast the enrollment with population in next decade. The result of forecasting for enrollment in next decade (2019 to 2028) is displayed in Table 5 and Figure 5. The future trend of enrollment will decrease in the case country from 1,211,146 in 2019 to 904,242 in 2028. The results may provide useful information for alerting related policy makers or institutional leaders. The trend provides messages for adjusting recruitment policy or enhancing their competitiveness in future.

Table 5. Forecasts of enrollment from 2019-2028

Year	Forecasts	Lower	Upper
2019	1211146	1177377	1244914
2020	1177045	1125444	1228647
2021	1142945	1072611	1213278
2022	1108844	1019444	1198245
2023	1074744	966117	1183371
2024	1040644	912703	1168584
2025	1006543	859236	1153850
2026	972443	805734	1139151
2027	938342	752209	1124476
2028	904242	698666	1109817

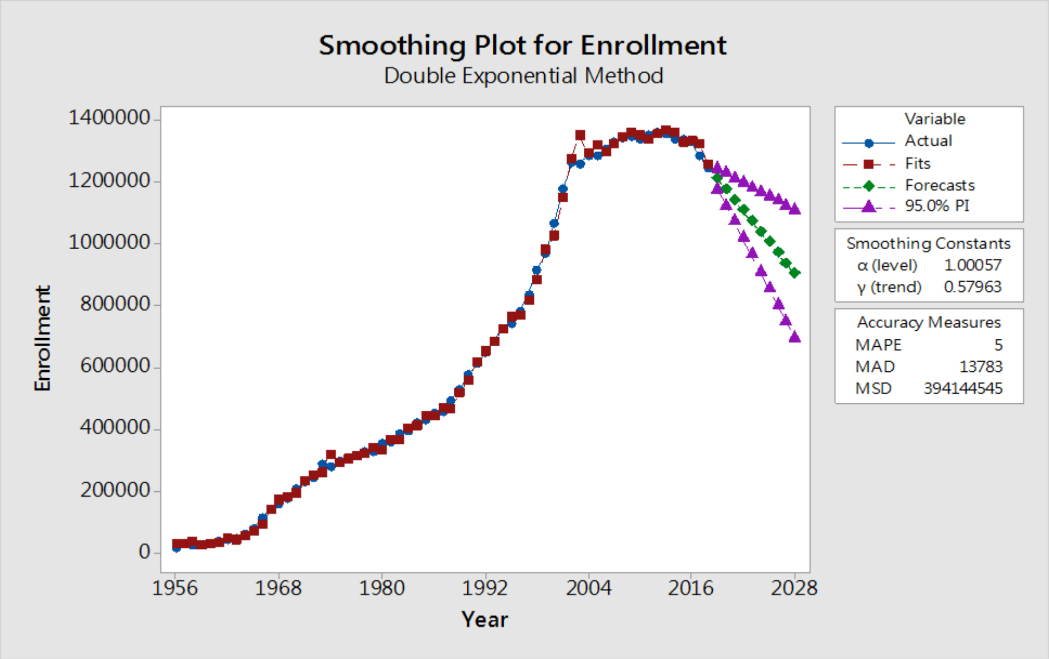

Figure 5. Result of predicted enrollment with ARIMAX$(2, 1, 1)$ model

Table 6. Compared the ARIMAX$(2, 1, 1)$ and ARIMA$(2, 1, 1)$ for enrollment

Models	ARIMAX$(2, 1, 1)$		ARIMA$(2, 1, 1)$	
Compared	Discrepancy (A-P)	Residual	Discrepancy (A-P)	Residual
Total error	4,196	-0.0612	$-46{,}139$	2
Average error	68.787		-744.177	

3.4. **Comparison of ARIMA and ARIMAX.** Table 6 displays the results of ARIMA$(2, 1, 1)$ for enrollment and ARIMAX$(2, 1, 1)$ for enrollment with population during the model building periods. We found the discrepancy (A-P) between actual values (A) and predicted values (P) in ARIMAX$(2, 1, 1)$ is smaller than that of ARIMA$(2, 1, 1)$. Noise residual in ARIMAX$(2, 1, 1)$ is -0.061, while in ARIMA$(2, 1, 1)$ is 2.

4. **Conclusions.** This study provides an example of ARIMAX model building to tackle time series data sets with their concurrent relationships. The selected ARIMAX$(2, 1, 1)$ model with population and enrollment can be used to predict enrollment in future. Considering the uncertain enrollment in higher education setting, the finding provides clear

trend that will decrease rapidly. Since the system has faced the new crisis of declining birthrate, the population will decrease steadily which might impact the enrollment directly. Moreover, the higher education expanding has shown a new high in the system for couple years ago. The findings may provide useful information for related policy makers to adjust related enrollment policy.

The ARIMAX will work well compared with that of ARIMA model in this case study. While the logic of series data transformation is a crucial component of model building. For further studies, this study suggests selecting fitted concurrent series data and using ARIMAX to tackle the similar issues in other settings. We suggest creating an innovative concept framework for related studies before your model building. The related statistical software can help the CCF test and ARIMAX model building.

REFERENCES

[1] 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.

[2] 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.

[3] C. Yuan, S. Liu and Z. Fang, Comparison of China's primary energy consumption forecasting by using ARIMA (the autoregressive integrated moving average) model and GM(1,1) model, *Energy*, no.100, pp.384-390, 2016.

[4] D.-F. Chang and H.-C. ChangTzeng, Patterns of gender parity in the humanities and STEM programs: The trajectory under the expanded higher education system, *Studies in Higher Education*, DOI: 10.1080/03075079.2018.1550479, 2018.

[5] Duke People, *Introduction to ARIMA Models*, https://people.duke.edu/∼rnau/411arim.htm, 2019.

[6] R. Nau, *Notes on Nonseasonal ARIMA Models*, Master Thesis, Fuqua School of Business, Duke University, 2019.

[7] C. Kongcharoen and T. Kruangpradit, Autoregressive integrated moving average with explanatory variable (ARIMAX) model for Thailand export, *The 33rd International Symposium on Forecasting*, Seoul, Korea, 2013.

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

[9] M. Yang, J. Xie, P. Mao, C. Wang and Z. Ye, Application of the ARIMAX model on forecasting freeway traffic flow, *The 17th COTA International Conference of Transportation Professionals*, 2017.

[10] C. Chatfield, *Time Series Forecasting*, Chapman and Hall, London, 2001.

[11] G. E. P. Box, G. M. Jenkins and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd Edition, Upper Saddle River, Prentice-Hall, NJ, 1994.

[12] J. D. Hamilton, *Time Series Forecasting*, Princeton University Press, Princeton, 1994.

[13] National Statistics, Taiwan, *National Income Statistics Information*, https://www.stat.gov.tw/ct.asp?xItem=37407&CtNode=3564&mp=4, 2019.

[14] Ministry of Education, *Statistics for Higher Education 2019*, https://stats.moe.gov.tw/bookcase/Higher/108/index.html#p=1, 2019.

[15] A. E. Usoro, Some basic properties of cross-correlation functions of n-dimensional vector time series, *Journal of Statistical and Econometric Methods*, vol.4, no.1, pp.63-71, 2015.

[16] K. V. Mardia and C. R. Goodall, Spartial-temporal analysis of multivariate environmental monitoring data, in *Multivariate Environmental Statistics, North Tolland Series in Statistics & Probability*, G. P. Patil and C. Radhakrishna Rao (eds.), NorthHolland, Amsterdam, 1993.

[17] E. Rublíková and L. Marek, Linear transfer function model for outflow rates, *Ekonomické rozhl'ady*, vol.30, no.4, pp.457-466, 2001.

[18] G. E. P. Box, G. M. Jenkins and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, John Wiley & Sons, New York, 2008.

[19] G. E. P. Box and G. C. Tiao, Intervention analysis with applications to economic and environmental problems, *Journal of the American Statistical Association*, vol.70, no.349, pp.70-79, 1975.

[20] Ď. Peter and P. Silvia, ARIMA vs. ARIMAX – Which approach is better to analyze and forecast macroeconomic time series?, *Proc. of the 30th International Conference Mathematical Methods in Economics*, pp.136-140, https://s3.amazonaws.com/academia.edu.documents/36551553/024_Durka.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1555666131&Signature=StnPO2m07NXYEjWhftTMhxUhZKg%3D&response-content-disposition=inline%3B%20filename%3DARIMA_vs._ARIMAX_which_approach_is_bette.pdf, 2012.