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

ICIC International ©2020 ISSN 2185-2766
pp. **129–136**

DETECTING MULTIVARIATE SERIES DATA WITH TRANSFER FUNCTION ARIMAX FOR TEACHER DEMAND

Wen-Ching Chou[1], Ko Lin Lai[2] and Dian-Fu Chang[1,*]

[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
{ sauichou1226; colin19750918 }@gmail.com
*Corresponding author: 140626@mail.tku.edu.tw

Received August 2019; accepted November 2019

Abstract. *The universal ARIMA model did not fit two series concurrently. For example, the series of teacher demand is a case with multiple factors concurrently in its development process. Previous literature assumed that teacher demand can be forecasted independently through the model ARIMA (autoregressive integrated moving average). In this paper, we applied multivariate ARIMA mode, called ARIMAX, for determining the demand of teachers with the series data of students. We selected the series of data as an example from the Ministry of Education, Taiwan. The cross correlation function and transfer function have been applied to build the fittest ARIMAX model. The ARIMAX analyses indicate that predicting the number of teachers with number of students is better than that only using the universal series data of teacher demand. The implication of these findings may provide an optimal research design for determining the demand of teachers under the changing number of students in elementary schools.*
Keywords: ARIMA, ARIMAX, Cross-correlation function, Multivariate time series, Transfer function, Teacher demand

1. **Introduction.** Time series analyses can be exposed by considering real experimental data taken from different subject areas [1]. For example, time series occur in the field of economics, where we are continually exposed to daily stock market quotations or monthly unemployment figures. An epidemiologist might be interested in the number of influenza cases observed over some time periods. In medicine, blood pressure measurements traced over time could be useful for evaluating drugs used in treating hypertension. Functional magnetic resonance imaging of brain-wave time series patterns might be used to study how the brain reacts to certain stimuli under various experimental conditions. Social scientists follow population series, such as birthrates or school enrollments [1]. The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data. Various time series studies have focused on ARIMA (autoregressive integrated moving average) models with series data [2-7], while g series data have become emerging topics in social science [8,9]. Compared with other fields, we found limited studies related to this topic in education setting.

Recently, balancing teacher supply and demand has become a hot topic in various education systems. For example, shortages of teachers in the United States are particularly severe in special education, mathematics, science, and bilingual/English learner education [10]. The United States' education system reflects the fact that teacher demand is growing, while the teacher supply is shrinking. In Australia, demand for teachers is largely a result of the number of children in the population. In most states, the primary school

DOI: 10.24507/icicelb.11.02.129

student population has been declining slowly since 2001 [11]. Green, Adendorff & Mathebula's study moves beyond simply basing an analysis of supply and demand on teacher attrition, and takes into account multiple variables that should be considered in supply and demand planning [12]. According to the data from the Ministry of the Interior, the newborn babies decreased from 328,461 in 1974 to 196,973 in 2016, a 40% drop [13]. The dramatically increasing drop in numbers has directly impacted the demand of elementary education teachers. This trend will impact on the participation in elementary education and cause oversupply of teachers. This study perceives that the situation is been going to worsen in the current setting. This study aims to develop a practical approach to work with multiple series in the proposed ARIMAX (multivariable autoregressive integrated moving average) model for ameliorating the issues of teacher demand. Given this purpose, this study will address the following research questions:

a) How to detect the series cross relationships to fit ARIMAX model for predicting teacher demand?

b) How to work with transfer function in the ARIMAX model with these series data?

c) How to select the fittest ARIMAX model for future teacher demand?

In this paper, we will conduct ARIMAX process with the selected series data sets. The research result will provide a fittest model to predict future teacher demand. The following section of this paper will begin with series investigation with cross correlation function (CCF) and transfer function to select ARIMAX models. Then, the example of series data with cross correlation will be addressed. Finally, the conclusion will be drawn.

2. **Method.**

We may consider a time series as a sequence of random variables, x_1, x_2, x_3, ..., where the random variable x_1 denotes the value taken by the series at the first time point, the variable x_2 denotes the value for the second time period, x_3 denotes the value for the third time period, and so on. In general, a collection of random variables, $\{x_t\}$, indexed by t is referred to as a *stochastic process*. In this text, t will typically be discrete and varies over the integers $t = 0, \pm 1, \pm 2, \ldots$, or some subset of the integers. The non-seasonal ARIMA(p, d, q) model is defined as [8]

$$\varphi(B)(1 - B)^d y_t = c + \theta(B)\varepsilon_t$$

where φ is the AR polynomial of order p and θ is the MA polynomial of order q, B is the backshift operator, $\{\varepsilon_t\}$ is a white noise process with zero mean and variance σ^2, and c is a constant.

A multivariate time series has more than one time-dependent variable. Each variable depends not only on its past values but also has some dependency on other variables. For calculating $y_1(t)$, we will use the past value of y_1 and y_2. Similarly, to calculate $y_2(t)$, past values of both y_1 and y_2 will be used. Below is a simple mathematical way of representing this relation [14,15]:

$$y_1(t) = a_1 + w_1 11 * y_1(t-1) + w_1 12 * y_2(t-1) + e_1(t-1)$$
$$y_2(t) = a_2 + w_1 21 * y_1(t-1) + w_1 22 * y_2(t-1) + e_2(t-1)$$

Here, a_1 and a_2 are the constant terms; $w_1 11$, $w_1 12$, $w_1 21$, and $w_1 22$ are the coefficients; e_1 and e_2 are the error terms.

2.1. **The process of conducting ARIMAX.**

To detect series with cross relationships, we track the following steps.

a) The first step in any time series investigation always involves careful examination of the recorded data plotted over time. This scrutiny often suggests the method of analysis as well as statistics that will be of use in summarizing the information in the data.

b) A simple kind of generated series might be a collection of uncorrelated random variables, w_t, with mean 0 and finite variance σ_w^2. The time series generated from uncorrelated variables is used as a model for noise in its applications, where it is called *white noise*.

The plot tends to show visually a mixture of many different kinds of oscillations in the white noise series.

c) CCF is used to detect the series with positive, negative or no correlation. The CCF generalizes the autocorrelation function (ACF) to the multivariate case. Thus, its main purpose is to find linear dynamic relationships in time series data that have been generated from stationary processes. The sample cross-correlation function (CCF) of the two series is defined as [16,17].

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

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 $c_{YY}(0)$ 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 SPSS (statistical package of social science), the program will provide cross correlation coefficient with ± 7 at least and CCF plot. When the series with cross correlation, the CCF will display significantly with positive or negative lags.

2.2. **Research target.** The major reason for selecting elementary education level in this study is because it has become an emerging issue for teacher supply under declining birthrate. Because of the low birth rate, the number of newborn babies has dropped from 410,000 in 1981 to 270,000 in 1998 and 167,000 in 2010, the lowest level in the last 50 years [18,19]. According to the data from the Ministry of the Interior, the newborn babies have dropped 40% [13]. While the number of students and number of teachers are relative series. To detect the real number of teachers, we found it is different from traditional ARIMA model with only one series.

2.3. **Technical programs in SPSS.** The SPSS programs for selected ARIMAX with *Num_teacher* (dependent variable) and *Num_student* (independent variable) are listed as follows.

```
TSMODEL
  /MODELSUMMARY PRINT = [MODELFIT RESIDACF RESIDPACF] PLOT = [SRSQUARE
RSQUARE RMSE MAPE MAE MAXAPE MAXAE NORMBIC RESIDACFRESIDPACF]
  /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(N = 1)] MODELFIT = NORMBIC
  /SAVE PREDICTED LCL(LCL) UCL(UCL) NRESIDUAL(NResidual)
  /AUXILIARY CILEVEL = 95 MAXACFLAGS = 24
  /MISSING USERMISSING = INCLUDE
  /MODEL DEPENDENT = Num_teacher INDEPENDENT = Num_student
      PREFIX = 'Model'
  /ARIMA AR = [1] DIFF = 2 MA = [1]
      TRANSFORM = LN CONSTANT = NO
  /TRANSFERFUNCTION VARIABLES = Num_student NUM = [1] DENOM = [2, 1]
DIFF = 1 DELAY = 0 TRANSFORM = LN
  /AUTOOUTLIER DETECT = OFF.
```

2.4. **Verification of the ARIMAX models.** The ARIMAX selection will follow the statistics, like smoothing R^2, R^2, RMSE, MAPE, MaxAPE, MAE, MaxAE, and standardized Bayesian information criterion (BIC). These statistics will display in the result of running SPSS program. In this study, we select the smaller standardized BIC as the proposed model. The formula for the BIC is listed as follows [20,21].

$$BIC = -2 * \ln(L) + k \ln(n)$$

n = sample size, k = the number of free parameters to be estimated, L = the maximized value of the likelihood function for the estimated model.

Under the assumption that the model errors or disturbances are normally distributed, this becomes

$$BIC = n \ln\left(\frac{RSS}{n}\right) + k \ln(n)$$

RSS = residual sum of squares from the estimated model.

Moreover, the SPSS will provide related estimations of parameters for the fittest ARIMAX model.

3. **Result.** In this section, we demonstrate how the two selected series have been tested with their cross correlation function for fit ARIMAX model with transfer function. Then, take the number of teachers and number of students as an example to conduct ARIMAX with transfer function by using SPSS. In the final section, the comparison of ARIMAX with transfer function and universal ARIMA model will be addressed.

3.1. **Testing with CCF.** In this study, we selected two series from 1972 to 2017 as examples. Table 1 displays they have strong cross correlations with number of students and number of teachers under one difference. The cross correlations are significantly demonstrated in lag 0 = 0.53, lag 1 = 0.44, lag 2 = 0.35, whereas lag −1 is 0.51, lag −2 is 0.40, and lag −3 is 0.36. The significances of CCF are displayed in Figure 1. The significant lags have shown from −4 to 3; it implies the two series could work well in the ARIMAX model with their negative tendency.

TABLE 1. Cross-correlation function with number of students and number of teachers (1972-2017)

Lag	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7
Cross-correlation	0.22	0.19	0.25	0.35	0.36	0.40	0.51	0.53	0.44	0.35	0.31	0.23	0.06	−0.01	0.04
Standard error	0.16	0.16	0.16	0.16	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.16	0.16	0.16	0.16

3.2. **Building fittest ARIMAX model with transfer function.** In this study, we selected ARIMAX(1,2,1) with natural logarithm the series data as the fittest model. The reason for conducting natural logarithm is that the two series data in terms of number of teachers and number of students are with wide discrepancies. When we deal with natural logarithm, it will smooth the trends of series for the predicted model. The results of ACF and PACF with SPSS are demonstrated in Figure 2. The time lags with ACF and PACF show there are no specific trends after two times differences with this series.

Based on the ARIMAX(1,2,1), the fittest statistics with mean and their values in different percentage for predicting the number of teachers are displayed in Table 2. The results reveal the mean of smooth R^2 = 0.315, R^2 = 0.994, RMSE = 1040.147, MAPE = 0.874, MaxAPE = 2.826, MAE = 761.317, MaxAE = 2485.334, and standardized BIC is 14.428. The Ljung-Box(18) equals 13.670 (df = 16, p = 0.623) which indicates the model meets the assumption that the residuals are independent. Moreover, the different percentages in this model also show the fittest values change very limited.

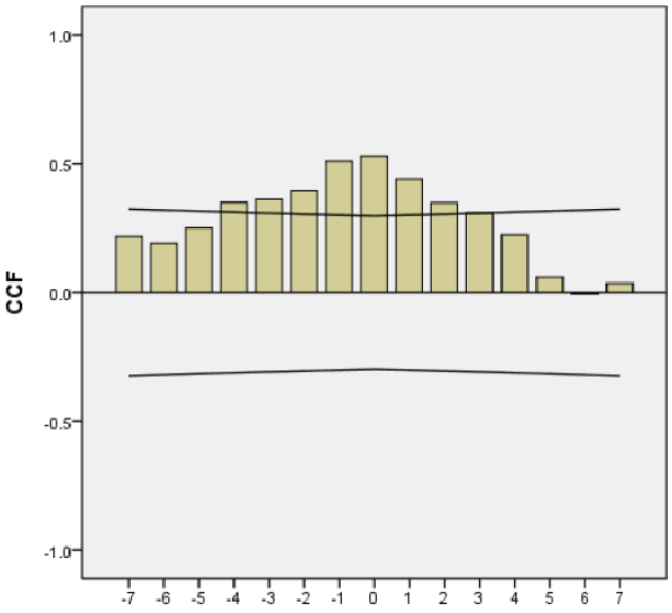

FIGURE 1. Testing the significance of CCF

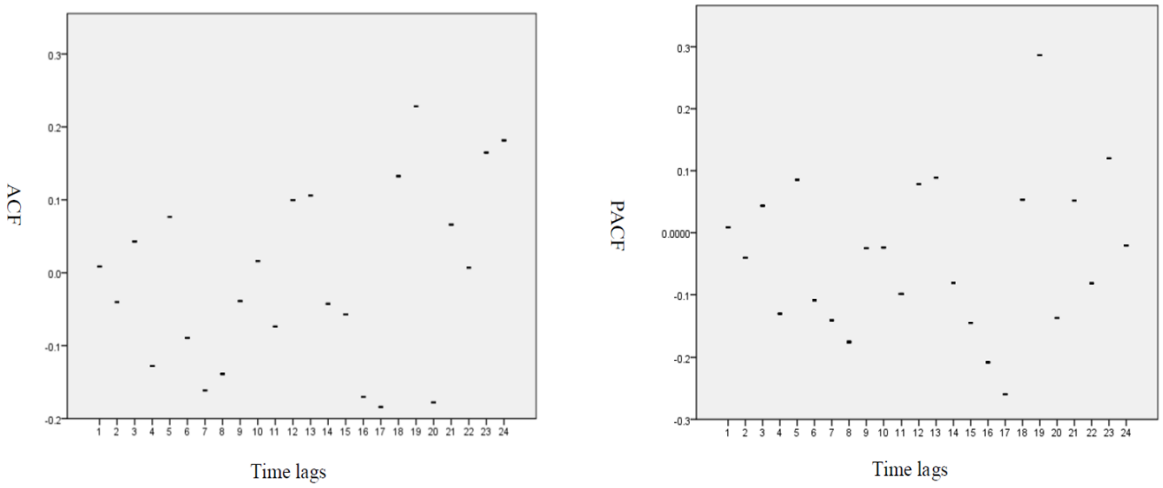

FIGURE 2. ACF and PACF for transfer function ARIMAX(1,2,1)

TABLE 2. The fittest statistics of ARIMAX(1,2,1) with transfer function

Fittest statistics	Mean	Percentage						
		5	10	25	50	75	90	95
Smooth R^2	.315	.315	.315	.315	.315	.315	.315	.315
R^2	.994	.994	.994	.994	.994	.994	.994	.994
RMSE	1040.147	1040.147	1040.147	1040.147	1040.147	1040.147	1040.147	1040.147
MAPE	.874	.874	.874	.874	.874	.874	.874	.874
MaxAPE	2.826	2.826	2.826	2.826	2.826	2.826	2.826	2.826
MAE	761.317	761.317	761.317	761.317	761.317	761.317	761.317	761.317
MaxAE	2485.334	2485.334	2485.334	2485.334	2485.334	2485.334	2485.334	2485.334
Std. BIC	14.428	14.428	14.428	14.428	14.428	14.428	14.428	14.428

The parameters of the ARIMAX(1,2,1) under natural logarithm are shown in Table 3. For predicting number of teachers, the fittest model shows AR(1) = 0.620, MA(1) = 0.994 with number of teachers when number of students as denominator is $-.961$ (lag = 1, $p = 0.000$) and $-.989$ (lag = 2, $p = .000$) with one time difference. When the number

of students is as numerator, the parameter is 0.339 (lag = 0, $p = .020$), while the other parameter equals 0.249 (lag = 1, $p = .097$) which did not fit to the .05 significant level. This model shows the gaps between observed and predicted values are the smallest. The visualized plot for the result is demonstrated in Figure 3.

TABLE 3. The fittest ARIMAX(1,2,1) with transfer function based on standardized BIC

Model					Estimate	SE	t	p
Num_teacher	Num_teacher	log	AR	Lag = 1	.620	.196	3.170	.003
			Difference = 2					
			MA	Lag = 1	.994	1.235	.805	.426
	Num_student	log	Numerator	Lag = 0	.339	.139	2.429	.020
				Lag = 1	.249	.147	1.702	.097
			Difference = 1					
			Denominator	Lag = 1	−.961	.041	−23.69	.000
				Lag = 2	−.989	.048	−20.69	.000

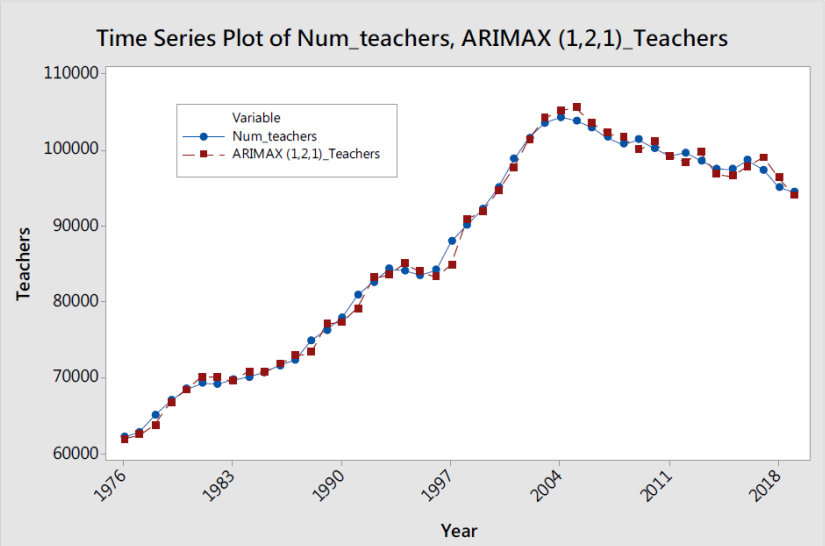

FIGURE 3. The visualized plot for teacher demand predicted values

3.3. **Comparing the transfer function ARIMAX with universal ARIMA.** In the fittest model selected process, we found two times difference of the series with transfer function model working well with the ARIMAX(1,2,1). When we check the universal ARIMA(1,2,1), it also fits with that function. Compared with the two models, we found the results are different. Table 4 shows the ARIMAX(1,2,1) with transfer function is better than that of universal ARIMA(1,2,1) due to their slim errors of prediction. Both models are similar to deal with the series of teacher demand, while the ARIMAX is superior to the ARIMA for predicting.

4. **Conclusion.** This study found long run relationship exists between the number of teachers and number of students confirming that there is a dependence between the variables. In general, researchers have considered each series independently. However, the current study shows that long run relationship between number of teachers and students should be considered concurrently. This study provides an example to tackle the issue

TABLE 4. Comparing the prediction of teacher demand with transfer function ARIMAX(1,2,1) and ARIMA(1,2,1)

Year	Transfer function ARIMAX(1,2,1)		ARIMA(1,2,1)		Year	Transfer function ARIMAX(1,2,1)		ARIMA(1,2,1)	
	$d=2$	Errors	$d=2$	Errors		$d=2$	Errors	$d=2$	Errors
1974			61815	293.84	1996	90294	0.00	90637	−509.78
1975			62608	194.89	1997	91888	0.00	91769	334.53
1976	63316	0.03	63383	1591.15	1998	94329	0.01	93605	1424.32
1977	67151	0.00	66701	252.73	1999	97350	0.01	97177	1568.14
1978	68739	−0.01	68573	−159.85	2000	101544	0.00	101442	138.91
1979	69459	0.00	69664	−457.49	2001	104143	−0.01	103700	−199.26
1980	70646	−0.02	69972	−830.65	2002	105414	−0.01	105008	−708.24
1981	69314	0.00	69264	348.59	2003	105047	−0.01	105047	−1244.22
1982	69858	0.00	70097	−41.89	2004	104237	−0.01	103662	−779.58
1983	71484	−0.01	70502	146.05	2005	102428	−0.01	102429	−767.39
1984	71133	0.01	71188	324.01	2006	100764	0.00	100980	−287.8
1985	72121	0.00	72233	53.68	2007	101149	0.00	100149	−1211.45
1986	74068	0.01	72940	1898.34	2008	101480	−0.01	101893	−1686.54
1987	76627	−0.01	76731	−505.35	2009	98637	0.01	99503	−347.78
1988	77180	0.01	77327	565.2	2010	99422	0.00	98495	1067.39
1989	80146	0.01	79187	1662.03	2011	99637	−0.01	99856	−1296.62
1990	82843	0.00	83045	−461.62	2012	97017	0.01	97895	−429.38
1991	83830	0.01	83953	350.9	2013	96864	0.01	96719	716.83
1992	86355	−0.03	85667	−1615.06	2014	97397	0.01	97378	1201.85
1993	83605	0.00	84057	−577.34	2015	98943	−0.02	99294	−1926.02
1994	83219	0.01	83240	909.69	2016	96358	−0.01	96502	−1420.9
1995	85449	0.03	84734	3200.16	2017	93396	0.01	93474	933.09
					Sum of errors:		−0.01		4135.01

to determine the demand of teachers in practice. The ARIMAX model with rigidly verification process can be used to predict the related time series data set. In forecasting, and even in economics or other fields, multivariate models are not necessarily better than univariate ones. While multivariate models are convenient in modeling interesting interdependencies and achieve a better (not worse) fit within a given sample. This study can be applied to solve similar issues in other settings. For further studies, we encourage selecting influential factors into the model to enhance the robust of prediction, for example, supply side of teachers and the numbers of retired teachers.

REFERENCES

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

[2] D.-F. Chang, B.-C. Sheng and W.-C. Chou, Two-stage approach for detecting teacher's supply and demand issues in elementary education, *ICIC Express Letters, Part B: Applications*, vol.10, no.4, pp.319-326, 2019.

[3] P. Rotela Jr., F. L. R. Salomon and E. de Oliveira Pamplona, ARIMA: An applied time series forecasting model for the Bovespa stock index, *Applied Mathematics*, no.5, pp.3383-3391, 2014.

[4] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, Springer, Netherlands, 2016.

[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 (the autoregressive integrated moving average) model and GM(1,1) model, *Energy*, no.100, pp.384-390, 2016.
[7] 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.
[8] 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, DOI: 10.1080/0735648X.2017.1360194, 2018.
[9] N. Achille, S. Haberman and G. Consigli, A multivariate approach to project common trends in mortality indices, pp.1-26, http://dx.doi.org/10.2139/ssrn.3149989, 2018.
[10] L. Sutcher L. Darling-Hammond and D. Carver-Thomas, *A Coming Crisis in Teaching? Teacher Supply, Demand, and Shortages in the U.S (research brief)*, Learning Policy Institute, Palo Alto, CA, https://learningpolicyinstitute.org/product/coming-crisis-teaching-brief, 2016.
[11] P. R. Weldon, The teacher workforce in Australia: Supply, demand and data issues, *Australian Council for Educational Research, Policy Insight*, no.2, pp.1-20, 2015.
[12] W. Green, M. Adendorff and B. Mathebula, Minding the gap? A national foundation phase teacher supply and demand analysis: 2012-2020, *South African Journal of Childhood Education*, vol.4, no.3, pp.1-23, 2014.
[13] Department of Statistics, Ministry of Education, *Summary of Preschool – By Public or Private*, https://depart.moe.edu.tw/ED4500/cp.aspx?n=1B58E0B736635285&s=D04C74553DB60CAD, 2018.
[14] R. J. Hyndman and G. Athanasopoulos, 8 ARIMA models, *OTexts*, https://tanthiamhuat.files.wordpress.com/2015/12/chapter-8-arima-models.pdf, 2015.
[15] A. Singh, *A Multivariate Time Series Guide to Forecasting and Modeling (with Python Codes)*, https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/, 2018.
[16] C. Chatfield, *The Analysis of Time Series: An Introduction*, Chapman and Hall, London, 1996.
[17] 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.
[18] D.-S. Chen, *Higher Education in Taiwan: The Crisis of Rapid Expansion*, http://www.isa-sociology.org/universities-in-crisis/?p=417, 2010.
[19] H.-T. Huang, D.-F. Chang and H.-J. Weng, Selecting managerial strategies by using fuzzy logics for kindergarten under the decreasing birth rate, *ICIC Express Letters, Part B, Applications*, vol.8, no.10, pp.1439-1447, 2017.
[20] G. Schwarz, Estimating the dimension of a model, *Annals of Statistics*, vol.6, no.2, pp.461-464, 1978.
[21] C. M. Hurvich and C.-L. Tsai, Regression and time series model selection in small samples, *Biometrika*, vol.76, pp.297-307, 1989.