ICIC Express Letters
Part B: Applications
Volume **10**, Number **2**, February **2019**

ICIC International ©2019 ISSN 2185-2766
pp. 105–112

MINING GENDER PARITY PATTERNS IN STEM BY USING ARIMA MODEL

Dian-Fu Chang[1] and Hui Hu[2,*]

[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; *Corresponding author: 806760095@s06.tku.edu.tw

Received July 2018; accepted October 2018

Abstract. *During 1976-2017, Taiwan's higher education system has marched from mass to post-mass higher education stage. After higher education expansion, does the gender parity in STEM be changed? The study transformed the data of male and female participation in STEM into gender parity index (Ds) in the higher education system. ARIMA model was used to predict the Ds in next decade. The ACF and PACF were used to verify the model fit and the Ljung-Box chi-square statistic was used to determine whether the model meets the assumptions that the residuals are independent. The result reveals the Ds will change significantly accompanied by the expansion of the STEM in next decade. This change may be driven by the open job-market in the society. The projected trend of gender parity in STEM provides useful information to detect the potential human resources. Finally, the process of model building can be applied other settings to dealing with the similar issues.*

Keywords: ARIMA, Gender parity, Higher education expansion, Higher education, STEM

1. **Introduction.** During the last three decades, the capacity of Taiwan's higher education has expanded rapidly. In the academic year 2015-2016, there were 158 HEIs, comprising 126 universities, 19 independent colleges, and 13 junior colleges [1]. The number of students attending universities increased from 0.564 million in 1990 to 1.339 million in 2014. The popularization of education has led to a rapid increase in student enrollment, although the figure has leveled off in the last decade. As higher education in Taiwan transforms from a mass system to a universal system, public concerns regarding the system's equality are being raised [2]. Especially, the expansion accompanied with the major enrollment increasing in STEM in terms of Science, Technology, Engineering, and Mathematics. This study wants to realize after higher education expansion, does the gender parity in STEM be changed?

Previous studies focused on gender parity issue in STEM programs (Science, Technology, Engineering, and Mathematics), for example, Kanny et al. reviewed 324 full texts spanning the past 40 years of scholarly literature and found that the gender gap in college level STEM remains a persistent issue despite increased efforts to understand and address women's unequal participation [3]. Bradley argued that gender differentiation has declined surprisingly little compared to decades ago. Females are more likely to graduate from programs in education, arts, humanities, social sciences, and law, while males are more likely to graduate from programs in natural sciences, mathematics, and engineering [4]. This study argued that the expansion of higher education might shift the landscape of male dominant in STEM programs. Taking the advantage of data mining applications,

DOI: 10.24507/icicelb.10.02.105

this study selects a specific expanded higher education system to verify the argumentation: higher education expansion might change the gender disparity in STEM programs.

Data mining can be used extensively to develop innovate solutions that address complex practical problems in today's society and to meet ever-increasing challenges [5]. According to Pruengkarn et al.'s argument, there are two main categories for data mining: predictive methods and descriptive methods [5]. This study conducted the predictive methods to explore the trend of gender parity gap by using ARIMA (autocorrelation integrated moving average) model in STEM programs in Taiwan. In 1976, Taiwan's GER (gross entrance rate) in higher education was greater than 15%, implying that the system was entering the stage of mass higher education. In 1999, the GER in HEIs (higher education institutions) exceeded 50%, indicating that Taiwan's system had become universal according to Trow's definition [6]. Based on the number of students in HEIs, the female GER in the 18-21-year age group was 11.81% in 1950, and increased to 53.17% in 1999. The female GER (40.39%) was obviously greater than that of the male GER (38.54%) in 1995 [1]. The system shows significant changes in its participation according to gender during the higher education expansion in the later stage of GER 15%-50% and the state of GER over 50%. The number of female students was greater than that of male students in 1997, 1998, 1999, 2003, and 2014, respectively [7].

The purposes of this study are as follows: a). to realize the effect of higher education expansion on gender parity in STEM; b). to project the future development in the system. The structure of this paper begins with addressing the current gender issue in STEM. Then, provide a brief description of the data transformation process. Section 3 is to examine the robustness of predictive models in this study. Finally, the conclusions are presented.

2. **Method.**

The STEM programs have been selected from relevant departments. In this study, the enrollment data in STEM programs are extracted from the data bank of Ministry of Education from 1950 to 2017 in Taiwan [1,8]. This study focused on the gender parity issue reflected on the male and female's participation in STEM. Discriminant coefficient (D) has been developed by using Becker's concept. To realize the effect of higher education expansion on gender parity in STEM, Becker's D coefficient has been calculated in this study [9]. To project the D in future, this study conducted ARIMA model to verify the workable trajectory in the higher education system.

2.1. **Evaluation of gender parity.**

The gender parity index (GPI) is an important measure for evaluating gender equality in higher education settings. To standardize the effects of the population structure of the appropriate age groups, the GPI of the GER for each level of education is used. The GER is the number of students enrolled at a given level of education, regardless of age, and expressed as a percentage of the population in the theoretical age group for the same level of education. The GPI in higher education is expressed as a ratio of the female GER to the male GER. The formula for calculating the GPI is $100 \times$ ([GER in higher education for females]/[GER in higher education for males]). A GPI equal to 1 indicates parity between females and males. In general, a value less than 1 indicates disparity in favor of males and a value greater than 1 indicates disparity in favor of females [10]. However, the current data set cannot support the calculation of GPI completely.

Becker's D is an alternative index for evaluating gender parity. Becker defined the economics of discrimination and proposed the concept of a coefficient of discrimination [9]. Although the definitions of discrimination vary depending on the field of study, the original definition might have equally concerned skilled people from specific demographic groups who were treated differently by the labor market. A method for equalizing skills is required to assess whether wage differences are inconsistent with observed productivity

[11]. Becker provided an alternative index to estimate parity, which can be applied for testing the gender parity in a specific level of education [7,9]. D is defined as follows:

$$D = (EM/EF) - 1$$

EM represents the education opportunities for males in STEM programs;

EF represents the education opportunities for females in STEM programs.

In this study, the Ds represent the discriminant coefficient in STEM programs.

2.2. **Projecting the Ds in STEM.** The projected Ds (in terms of D in STEM) have been calculated from 2018 to 2027 with a selected ARIMA(p, d, q) model. This study follows the ARIMA model building process preparing data to obtain stationary series identify potential models, check ACF/PACF of residuals, and forecasting [12-15]. The process of ARIMA building stage has displayed in Figure 1 [13].

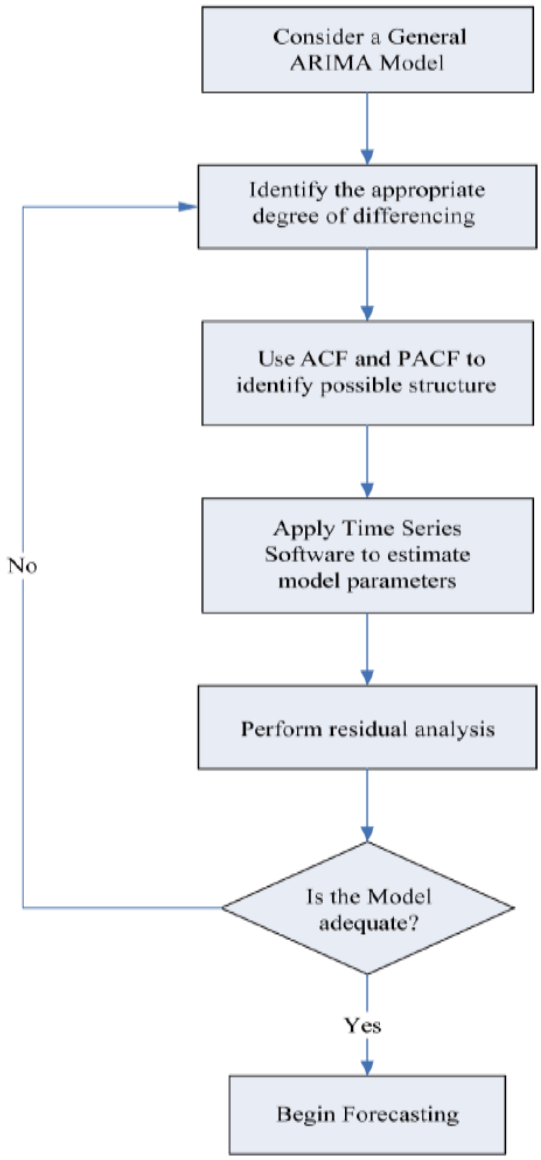

FIGURE 1. ARIMA building stages

A non-seasonal ARIMA model is classified as an "ARIMA(p, d, q)" model, where:
p is the number of autoregressive terms (AR),
d is the number of non-seasonal differences needed for stationarity (Difference), and
q is the number of lagged forecast errors in the prediction equation (MA).

The forecasting equation is constructed as follows. First, let y denote the d^{th} difference of Y, which means:

If $d = 0$: $y_t = Y_t$

If $d = 1$: $y_t = Y_t - Y_{t-1}$

If $d = 2$: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

Typically, differencing removes the linear trend. The differenced series provide more stationary type for forecasting. Initially, inspection of the ACF of a time series is necessary to determine if the series is stationary or will require differencing [13,14]. This study plots the ACF and PACF of the differenced series to look for the data more consistent with a stationary process to fit the model meets the assumptions that the residuals are independent. The analysis is carried out using the Minitab® statistical package [16]. This study used the Ljung-Box chi-square statistics to determine whether the model meets the assumptions that the residuals are independent [17]. To determine whether the residuals are independent, this study compared the p-value to the significance level for each chi-square statistic. The calculations are listed as follows [17,18]:

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

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

Usually, a significant level of .05 (denoted as α) works well. Basically, the p-values for the Ljung-Box chi-square statistics are all greater than .05.

3. **Results.**

3.1. **STEM trend of male and female students.**

The trend of male and female in STEM programs has shown in Figure 2. The largest gap has displayed in 2007. The system shows significant changes in its participation according to gender during the higher education expansion in the later stage of GER 15%-50% and the state of GER over 50%. The GER in this system increased from 15% (1976) to 50% (1999) during these 23 years. While the GER up to 85% happened in 2007, the system only spent another 8 years to reach the ceiling of GER [7]. The result reveals the higher education expansion has increased the gender parity gap in STEM directly in this system.

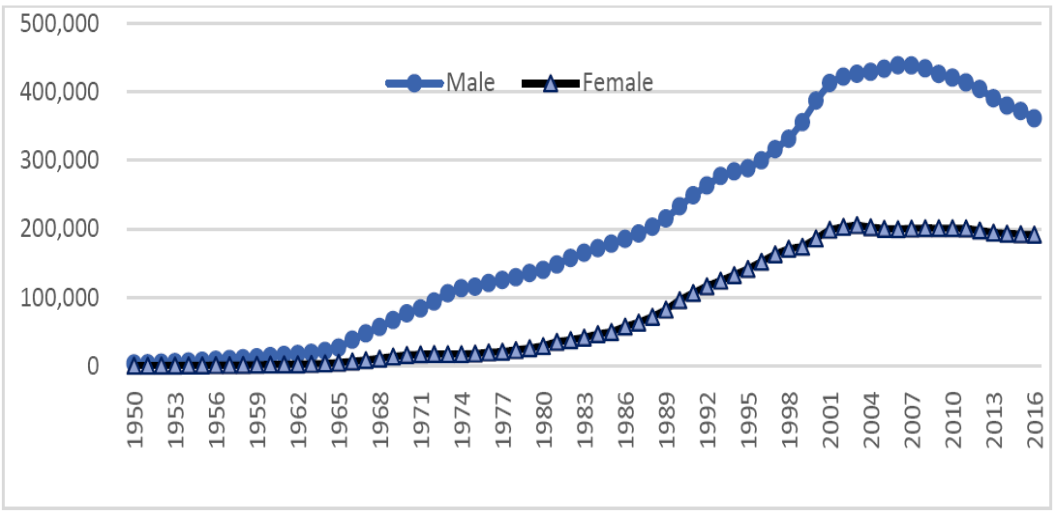

FIGURE 2. Males and females in STEM from 1950 to 2016

3.2. **Building time series model with D in STEM.** To realize the future trend of gender parity in STEM, this study employed ARIMA to build a predicated model for Ds in the next decade. Initially, inspection of the ACF of a time series is necessary to determine if the series is stationary or will require differencing. Because the data is not stationary, the series presents variance from one period to another needs to be different for station. The analysis used in this series demonstrated the need for a logarithmic transformation on the data. Tests for ACF (autocorrelation) and PACF (partial autocorrelation) indicated that the AR(1) model ARIMA$(1, 1, 0)$ model could be used to predict the behavior of the series, shown in Table 1. Evidence therefore exists to support that the residuals follow a white noise process and the AR(1) is a robust representation of the observed time series. The AR(1) process is defined by $y_t = -0.0649 + 0.8954y_{t-1}$. Actually, constant term's p-value is .340, and it implies there is no significant difference. The ACF and PACF diagrams of the residual values are returned in Figure 3 and do not show any discernible pattern. A verification of the series' residuals tests was carried out in this study. It demonstrated that autocorrelation does not exist between series residuals, which enabled the utilization to forecast the series, see Figure 4.

TABLE 1. Final estimation of parameters

Type	Coef.	SE Coef.	t-value	p-value
AR(1)	0.8954	0.0780	11.48	0.000
Constant	−0.0649	0.0674	−0.96	0.340

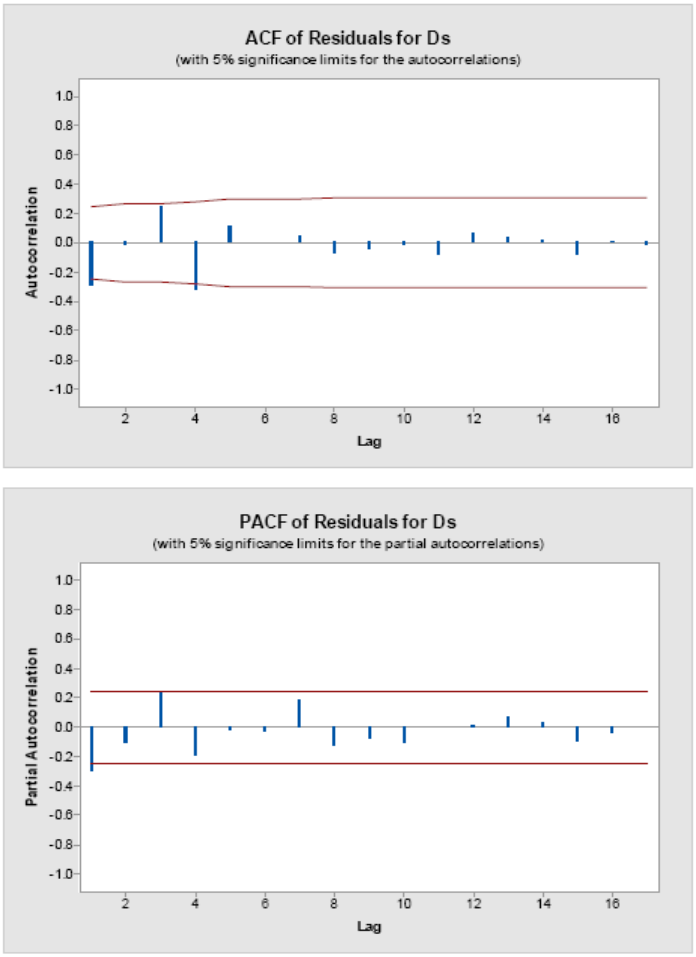

FIGURE 3. ACF and PACF for residuals of the ARIMA process

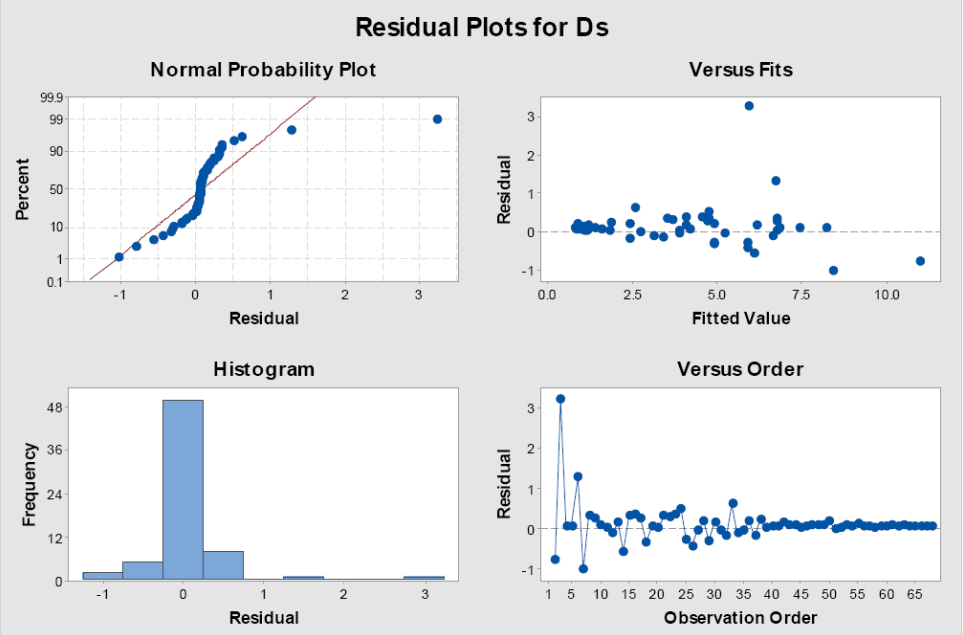

FIGURE 4. Residual plotting for predicting Ds

TABLE 2. Modified Box-Pierce (Ljung-Box) chi-square statistics

Lag	12	24	36	48
Chi-square	20.06	23.79	32.80	33.24
DF	10	22	34	46
p-value	0.029	0.358	0.526	0.920

Ljung-Box chi-square statistics demonstrate the model meets the assumptions that the residuals are independent. Typically, a significant level of .05 (denoted as α) works well. In this study, most of the p-values for the Ljung-Box chi-square statistics are all greater than .05, and only the lag 12 group shows $p = .029$. While the $p > .01$ means the problem with residuals at nonseasonal level still at an acceptable level, see Table 2.

3.3. **Forecasts from period 69 (2018) to 78 (2027).** The forecasting 10 years ahead in series of Ds demonstrates in Figure 5 and Table 3. This finding could be valid, if the prediction is used to interpret the gender parity trend in STEM with expansion system. The gaps between males and females in STEM are tended to diminish with the increasing expansion in the higher education system. The prediction of a decade ahead would determine actions to be taken by the related policy makers. However, it could hardly change the trend in the short term.

4. **Conclusions.** The study demonstrates the ARIMA$(1, 1, 0)$ model is acceptable to predict the Ds in STEM in the higher education setting. The statistics of the AR(1) model coefficients and chi-square statistics for modified Box-Pierce (Ljung-Box) provide proof to this prediction. The result reveals that the trend of Ds in current STEM system is declining, and the fact is that male still dominated the field. While accompanying the higher education expansion, it might change the trend of Ds in STEM in future. At this point, the result of ARIMA provides meaningful information for policy makers.

In further studies, how to take consideration of the technology driven in the open job-market may provide more clear picture of the development in the system. The study found ARIMA provides alternative options for selecting the fittest model to project the future trend. For further study, how to select different stages of expansion to test the

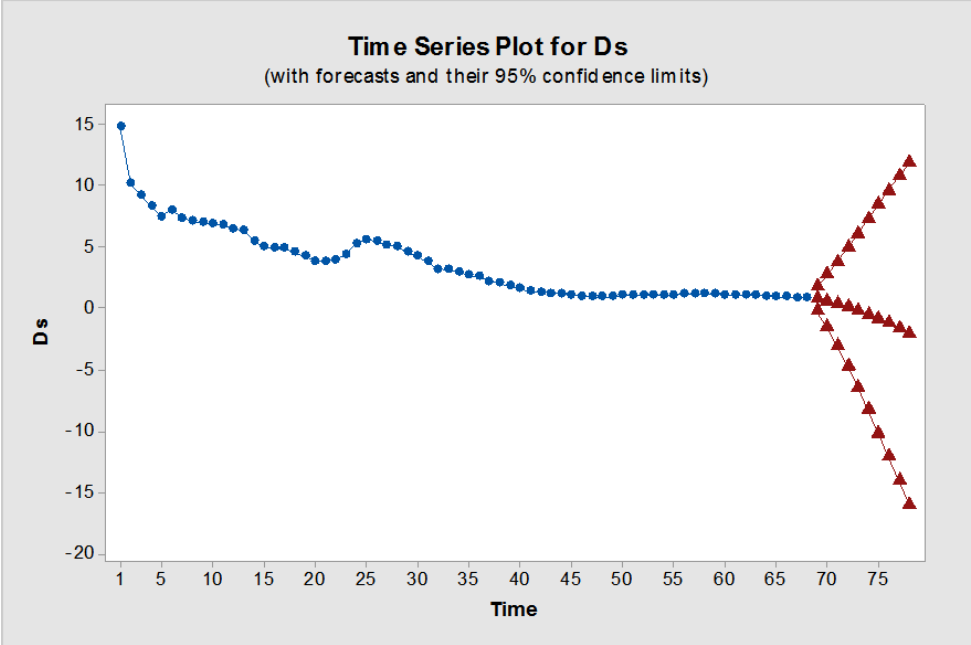

FIGURE 5. Time series plot for Ds in STEM from 1950 to 2027

TABLE 3. The prediction of Ds in STEM from 2018 to 2027

Period	Year	Forecast	95% limits	
			Lower	Upper
69	2018	0.73186	−0.2578	1.7215
70	2019	0.57123	−1.5496	2.6921
71	2020	0.36253	−3.0467	3.7718
72	2021	0.11080	−4.6898	4.9114
73	2022	−0.17947	−6.4384	6.0794
74	2023	−0.50425	−8.2633	7.2548
75	2024	−0.85992	−10.1426	8.4228
76	2025	−1.24325	−12.0599	9.5734
77	2026	−1.65136	−14.0020	10.6993
78	2027	−2.08166	−15.9589	11.7956

trend of Ds in STEM is important. The series in different expanded stages may provide more useful information for projecting Ds in future. Furthermore, the research design provides an example to verify the similar issues in different higher education settings.

REFERENCES

[1] Ministry of Education, *Summary of Tertiary Education Institutes (2016-2017)*, http://stats.moe.gov.tw/files/main_statistics/u.xls, 2017.

[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, pp.251-263, 2015.

[3] M. A. Kanny, L. J. Sax and T. A. Riggers-Piehl, Investigating forty years of STEM research: How explanations for the gender gap have evolved over time, *Journal of Women and Minorities in Science and Engineering*, vol.20, no.2, pp.127-148, 2014.

[4] K. Bradley, The incorporation of women into higher education: Paradoxical outcomes?, *Sociology of Education*, vol.73, no.1, pp.1-18, 2000.

[5] R. Pruengkarn, K. W. Wong and C. C. Fung, A review of data mining techniques and applications, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol.21, no.1, pp.32-48, 2017.

[6] M. Trow, Problems in the transition from elite to mass higher education, *ERIC*, ED 091983, http://files.eric.ed.gov/fulltext/ED091983.pdf, 1973.
[7] 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.
[8] Ministry of Education, *Predicting the Enrollment of Higher Education*, http://www.edu.tw/News_Plan_Content.aspx?n=D33B55D537402BAA&sms=954974C68391B710&s=889A5EA64C666473, 2018.
[9] G. S. Becker, *The Economics of Discrimination*, 2nd Edition, The University of Chicago Press, Chicago, 1971.
[10] The World Bank, *The Gross Enrolment Ratio, Tertiary, the Gender Parity Index (GPI)*, http://data.worldbank.org/indicator/SE.ENR.TERT.FM.ZS, 2016.
[11] P. Americana, *Becker's Taste for Discrimination*, http://praxamericana.blogspot.tw/2013/04/beckers-taste-for-discrimination.html, 2013.
[12] G. E. P. Box, G. M. Jenkins and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 4th Edition, Wiley & Sons, New York, 2008.
[13] S. Bisgard and M. Kulachi, *Time Series Analysis and Forecasting by Example*, Wiley & Sons, New York, 2011.
[14] R. Davies, T. Coole and D. Osipyw, The application of time series modelling and Monte Carlo simulation: Forecasting volatile inventory requirements, *Applied Mathematics*, no.5, pp.1152-1168, 2014.
[15] P. Rotela Jr., F. L. R. Salomon and E. de O. Pamplona, ARIMA: An applied time series forecasting model for the Bovespa stock index, *Applied Mathematics*, no.5, pp.3383-3391, 2014.
[16] Minitab, *Interpret the Key Results for ARIMA*, https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/arima/interpret-the-results/key-results/?SID=117600, 2018.
[17] G. Ljung and G. Box, On a measure of lack of fit in time series models, *Biometrika*, no.65, pp.297-303, 1978.
[18] R. Nau, *ARIMA Models for Time Series Forecasting*, https://people.duke.edu /∼rnau/411arim.htm, 2014.