Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
EMS015-1846

ARIMA by Box Jenkins Methodology for Estimation and Forecasting Models in Higher Education

Marilena Aura Din
Associate Professor
Romanian American University of Buchares
Romania

An Introduction to ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. This paper has been peer reviewed by at least two academic members of ATINER.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Din, M.A. (2016). **"ARIMA by Box Jenkins Methodology for Estimation and Forecasting Models in Higher Education",** Athens: ATINER'S Conference Paper Series, No: **EMS2015-1846.**

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209 Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm
Printed in Athens, Greece by the Athens Institute for Education and Research. All rights reserved. Reproduction is allowed for non-commercial purposes if the source is fully acknowledged.
ISSN: 2241-2891
09/03/2016

ARIMA by Box Jenkins Methodology for Estimation and Forecasting Models in Higher Education

Marilena Aura Din
Associate Professor
Romanian American University of Buchares
Romania

Abstract

Although there are many approaches in the existing literature which have attempted to treat education enrollments with specific forecasting models such as moving average, exponential smoothing, Markov chain, regression, Fuzzy time series, and others, the results were not enough to understand the moving data evolution, or not enough accuracy. Knowing that the Auto Regressive Integrated Moving Average (ARIMA) is one of the most powerful approaches to forecasting, mostly used to financial time series, this paper aims to see if data on students' enrollment in higher education can be adjusted to ARIMA models for estimation and forecasting purposes. The paper analyzes data from higher education enrollments within the ARIMA framework given by Box Jenkins Methodology. The empirical study revealed the best ARIMA validated model to be used to forecast future values for the next eight years. Even though such approach generally helps understanding data or predicts future points, after the steps of identification, estimation and verification followed here to build the best ARIMA model, the findings are not providing the foresight of the causes that may influence what will happen in the future years.

Keywords: Box-Jenkins methodology, ARIMA models, higher education enrollments forecasts.

Introduction

In higher education sector, the enrollment forecasting provides information to decision making and budget planning. The senior forecaster McCalman (McCalman 2012) gives three reasons for which decision makers need forecasts: ”the state of the world in the future is unknown; the success or failure of a decision made today will depend on the future; and decisions cannot be corrected afterwards (at least not without cost)”. Of course, the data limitations may restrict the decision makers to one or the other forecasting method. So that, in some cases, the higher education institutions are planning processes both a short term, ratio-based enrollment forecast as well as long-term, time series-based enrollment forecast.

Knowing that the Auto Regressive Integrated Moving Average (ARIMA) is one of the most powerful approaches for forecasting, mostly used to financial time series, this paper aims to see if data on student enrollment can be adjusted to ARIMA models for estimation and forecasting purposes.

The paper analyzes data from higher education enrollments within the ARIMA framework given by Box Jenkins Methodology. The data analyzed in this paper represent the percentage of the population 18 and 19 years old enrolled in higher education in United States, for the selected years 1963 through 2012. The source of data is the Digest of Education Statistics which includes a selection of data from many sources, both government[1] and private, and draws especially on the results of surveys and activities carried out by the National Center for Education Statistics (NCES).

There are many approaches to forecasting enrollments in education (Zhang 2007), such as: moving average, exponential smoothing, Markov chain (Sullivan and Woodall 1994), the traditional regression, Fuzzy time series (Chen and Hsu 2004) and a more complete list may continue as can be seen in the reference (Chen 2008). The moving average approach use projected values for the next years beyond historical data and this might result in large errors. The exponential smoothing method forecast one point at a time, adjusting its forecast as new data come in. The traditional regression is not advisable when the violation of the assumption of uncorrelated errors occur. Concerning reunification in Germany, Boes and Pflaumer (2005) analyzed the university student enrollment forecasts with outliers by structural ratios, using ARIMA-methods. Chen and Hsu (2004) proposed a Fuzzy time series method that belongs to first order and time-variant methods for a higher forecasting accuracy rate than the previous methods and apply it to forecasting enrollments.

In this paper we use fifty annually observations data, from 1963 to 2012, to model the time series of the percentage of the population 18 and 19 years old enrolled in higher education, noted here HEENROLL in order to identify a forecast model, estimate its parameters, check the model’s performance, and

[1]United States Census Bureau; CPS Historical Time Series Tables on School Enrollment, and Current Population Survey (CPS), October, 1970 through 2012

finally use it to forecast. The data have the advantage that they are being reasonable percentage values instead of nominal values, but the interpretation of the results will also be reported to the according population. Even though it is about fifty annually observation data, the length of the series could be quite small for the purpose. This limitation of students' enrollment data is a characteristic of the most countries.

An Autoregressive Integrated Moving Average (ARIMA) model is developed here using the Box-Jenkin's methodology in order to fit the best ARIMA model to the previous HEENREOLL time series. This model has to support the forecast of future values of HEENROLL. The empirical study revealed the best ARIMA validated model to forecast future values for the next eight years.

Auto Regressive Integrated Moving Average (Arima)

The ARIMA model has three parts: 1) the autoregressive part is a linear regression that relates past values of data series to future values 2) the integrated part indicates how many times the data series has to be differenced to get a stationary series, and 3) the moving average part that relates past forecast errors to future values of data series. The processes type are AR(p), MA(q), ARMA(p,q), ARIMA (p,d,q). The questions in mind we have are: how does one know whether it follows a purely AR process or a purely MA process or an ARMA process or an ARIMA process.

$$AR(p):\ x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t \quad MA(q):\ x_t = \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

$$AR(1):\ x_t = \phi x_{t-1} + \varepsilon_t \quad \text{which is a random walk for } \phi = 1$$

$MA(1):\ x_t = \theta\varepsilon_{t-1} + \varepsilon_t$, where the residuals $\{\varepsilon_t\}$ are White Noise distributed: $\varepsilon_t \sim i.i.d.N(0,\sigma^2)$. The stochastic process $\{\varepsilon_t\}$ is called White Noise (WN), if at every moment, the random variable ε_t is normally distributed, with zero mean and constant variance, i.e. meet the conditions:

$$E(\varepsilon_t) = 0;\ \mathrm{E}(\varepsilon_t^2) = \sigma^2;\ \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+k}) = 0,\ t \neq k$$

When using the delay operator (lag) $L(x_t) = x_{t-1}$, the processes can written:

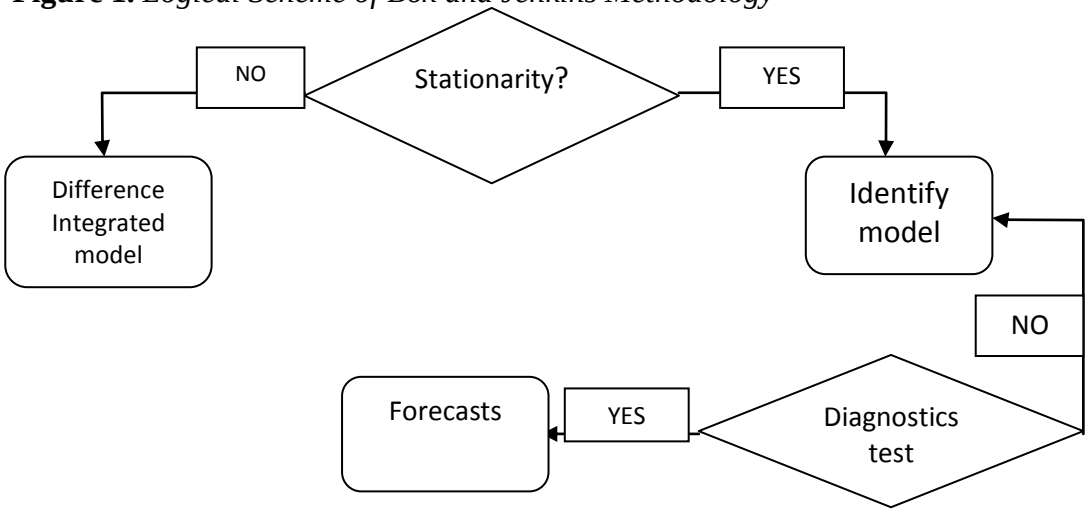
$$AR(p):\ \left(1 - \sum_{i=1}^{p} \phi_i L^t\right) x_t = \varepsilon_t,\ \text{i.e. } \Phi(L)x_t = \varepsilon_t,\ AR(1):\ (1-\phi L)x_t = \varepsilon_t$$

$$MA(q):\ x_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t,\ \text{i.e. } x_t = \Theta(L)\varepsilon_t,\ MA(1):\ x_t = (1+\theta L)\varepsilon_t$$

Methodological Approach for "Heenroll" Series of Data

The Box Jenkins (BJ) methodology (Johnston & DiNardo, 1997) is an iterative process, as in the figure 1. To use the Box-Jenkins methodology (Box-Jenkins 1976), we must have either a stationary time series or a time series that is stationary after one or more differencing.

Figure 1. *Logical Scheme of Box and Jenkins Methodology*

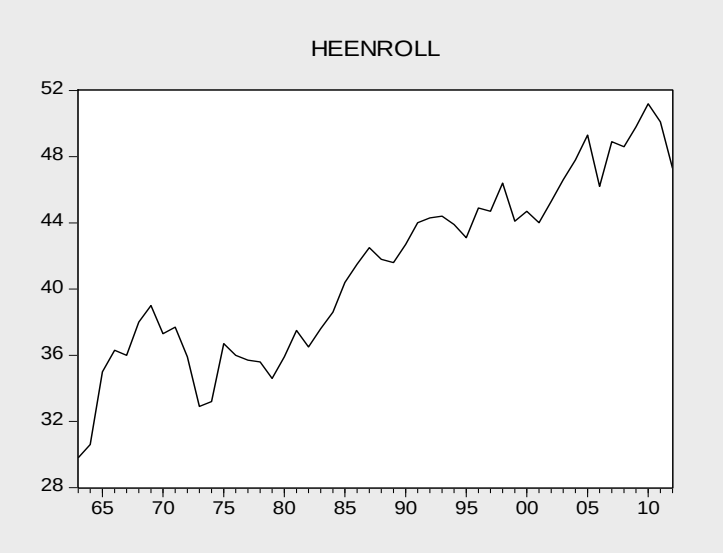


Step 1: Stationarity of the Data

The reason for assuming stationarity is to provide a valid basis for forecasting. A stochastic process is called stationary in the broad sense if it satisfies:

$$E\left(x_t\right)=\mu;\ \mathrm{E}\left[\left(x_t-\mu\right)^2\right]=\sigma_x^2;\ \mathrm{Cov}\left(x_t,x_{t+k}\right)=\gamma\left(k\right)$$

Figure 2. *HEENROLL Graph*



The first inspection of data from the plot of HEENROLL on time (Figure 2) suggests that the process is not stationary and that it is necessary to differentiate it. We study the stationarity of the data series HEENROLL by applying two tests where we check the null hypothesis H_0: "*series is not stationary*": a) unit root test Dickey-Fuller Augumented (ADF) in figure 3, and b) Phillips-Perron unit root test (PP) in figure 4.

Figure 3. *Augmented Dickey-Fuller Unit Root Test on HEENROLL*

Null Hypothesis: HEENROLL has a unit root

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.803694	0.3745
Test critical values:	1% level	-3.571310	
	5% level	-2.922449	
	10% level	-2.599224	

Figure 4. *Phillips-Perron Unit Root Test on HEENROLL*

Null Hypothesis: HEENROLL has a unit root

		Adj. t-Stat	Prob.*
Phillips-Perron test statistic		-1.802055	0.3753
Test critical values:	1% level	-3.571310	
	5% level	-2.922449	
	10% level	-2.599224	

In case of ADF test we have p-value higher than the significance level 0.05%, so we can not reject the hypothesis that the series is not stationary. According to PP tests from figure 4, we have the same conclusion, that the HEENROLL series of data is not stationary. The Autocorrelation Function (ACF) and the Partial Auto Correlation Function (PACF) also confirmed the HEENROLL series is not stationary, so we think to apply some transformations, such as logarithms and differentiating data series to ensure that the assumption of stationary of the ARIMA model is satisfied.

$$HEENROLL\_LN = LOG(HEENROLL)$$

Because we checked again the stationary condition for the new transformed data series, and as it is still not stationary, we differentiate it. The following transformation is named the difference of first order of the time series HEENROLL_LN, and is noted HEENROLL_LN_D.

$$D(HEENROLL\_LN) = HEENROLL\_LN - HEENROLL\_LN(-1)$$

Figure 5. *HEENROLL_LN_D graph*

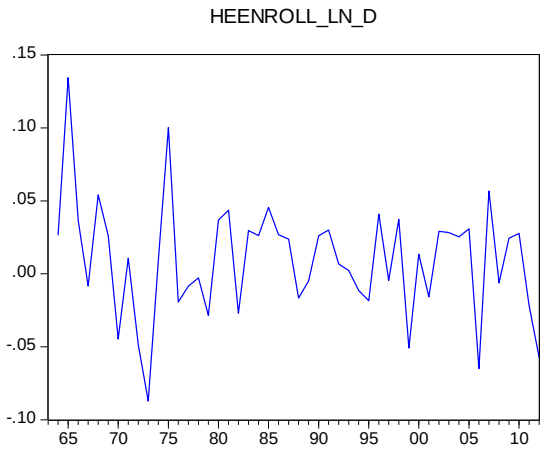


Figure 6. *Augmented Dickey-Fuller Test Statistic for D (HEENROLL_LN)*

Null Hypothesis: HEENROLL_LN_D has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic based on SIC, MAXLAG=10)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.476943	0.0000
Test critical values:	1% level	-3.574446	
	5% level	-2.923780	
	10% level	-2.599925	

Finally, the first difference of HEENROLL_LN turns out to be a stationary process, because the Augmented Dickey Fuller test confirms that we can reject the null hypothesis that the first difference HEENROLL_LN_D has a unit root (non-stationary) at the 5% significance level.

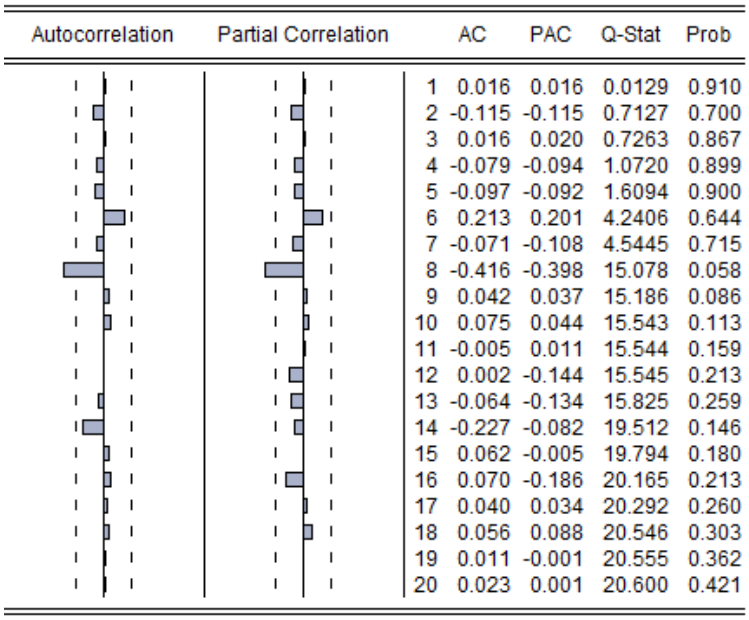
The next stage is to determine the p, d and q in the ARIMA (p, d, q) model.

Step 2: Identification/Specification through Correlogram: ACF and PACF

This step is to find out the appropriate values of p, d, and q using *correlogram* and *partial correlogram* and *Augmented Dickey Fuller Test*.
After we have estimated more models, the correlogram allows us to determine the possible candidates ARIMA (p,d,q) models. One of the candidates is the ARIMA (1;1;1) model:
HEENROLL_LN_D = C(1) + C(2)*HEENROLL_LN_D(-1) + C(3)*MA(1)

Figure 7. *Correlogram on HEENROLL_LN_D*

Sample: 1963 2012
Included observations: 49



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.016	0.016	0.0129	0.910
		2	-0.115	-0.115	0.7127	0.700
		3	0.016	0.020	0.7263	0.867
		4	-0.079	-0.094	1.0720	0.899
		5	-0.097	-0.092	1.6094	0.900
		6	0.213	0.201	4.2406	0.644
		7	-0.071	-0.108	4.5445	0.715
		8	-0.416	-0.398	15.078	0.058
		9	0.042	0.037	15.186	0.086
		10	0.075	0.044	15.543	0.113
		11	-0.005	0.011	15.544	0.159
		12	0.002	-0.144	15.545	0.213
		13	-0.064	-0.134	15.825	0.259
		14	-0.227	-0.082	19.512	0.146
		15	0.062	-0.005	19.794	0.180
		16	0.070	-0.186	20.165	0.213
		17	0.040	0.034	20.292	0.260
		18	0.056	0.088	20.546	0.303
		19	0.011	-0.001	20.555	0.362
		20	0.023	0.001	20.600	0.421

Step 3: Estimation (Estimation of Equation, Estimation of Coefficients)

This step is to estimate the parameters of the autoregressive and moving average terms included in the model (simple least squares or nonlinear in parameter estimation methods). The estimation is handled here by statistical package EViews (Johnson n.d.).

Figure 8. *Equation Estimation of ARIMA(1,1,1) model*

Dependent Variable: HEENROLL_LN_D
Method: Least Squares
Sample (adjusted): 1965 2012
Included observations: 48 after adjustments
Convergence achieved after 18 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.003043	0.002083	1.460860	0.1510
HEENROLL_LN_D(-1)	0.520764	0.228868	2.275394	0.0277
MA(1)	-0.775724	0.188332	-4.118916	0.0002

R-squared	0.183924	Mean dependent var	0.009073
Adjusted R-squared	0.147654	S.D. dependent var	0.040184
S.E. of regression	0.037099	Akaike info criterion	-3.689973
Sum squared resid	0.061936	Schwarz criterion	-3.573023
Log likelihood	91.55936	Hannan-Quinn criter.	-3.645778
F-statistic	5.070965	Durbin-Watson stat	1.615521
Prob(F-statistic)	0.010326		

Step 4: Diagnostic Checking

At this step we check that model is fit to the data, obtain residual, obtain ACF and PACF of residual, and apply different tests for diagnostic in order to validate the models and then to chose the best of them.

One simple test of the chosen model is to *see if the residuals estimated from this model are white noise;* if they are, we can accept the particular fit; if not, there is evidence of autocorrelation of errors, we need to go back to the identification stage and re-specify the model, by adding more lags.

Figure 9. *Actual, Fitted and Residual Graph*

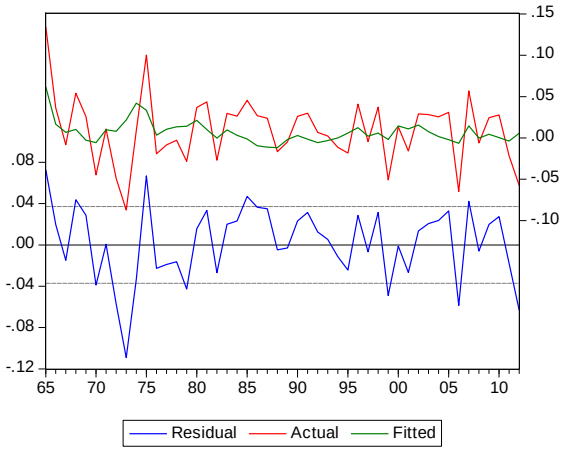
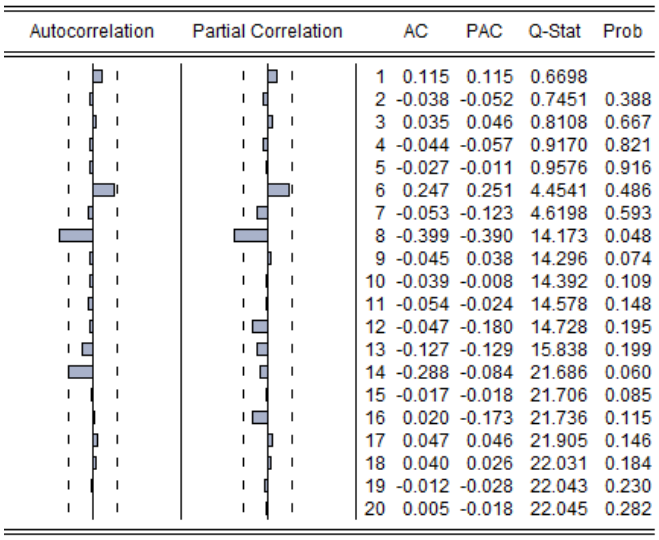


Figure 10. *Correlogram of Residuals*

Sample: 1965 2012
Included observations: 48
Q-statistic probabilities adjusted for 1 ARMA term(s)



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.115	0.115	0.6698	
		2	-0.038	-0.052	0.7451	0.388
		3	0.035	0.046	0.8108	0.667
		4	-0.044	-0.057	0.9170	0.821
		5	-0.027	-0.011	0.9576	0.916
		6	0.247	0.251	4.4541	0.486
		7	-0.053	-0.123	4.6198	0.593
		8	-0.399	-0.390	14.173	0.048
		9	-0.045	0.038	14.296	0.074
		10	-0.039	-0.008	14.392	0.109
		11	-0.054	-0.024	14.578	0.148
		12	-0.047	-0.180	14.728	0.195
		13	-0.127	-0.129	15.838	0.199
		14	-0.288	-0.084	21.686	0.060
		15	-0.017	-0.018	21.706	0.085
		16	0.020	-0.173	21.736	0.115
		17	0.047	0.046	21.905	0.146
		18	0.040	0.026	22.031	0.184
		19	-0.012	-0.028	22.043	0.230
		20	0.005	-0.018	22.045	0.282

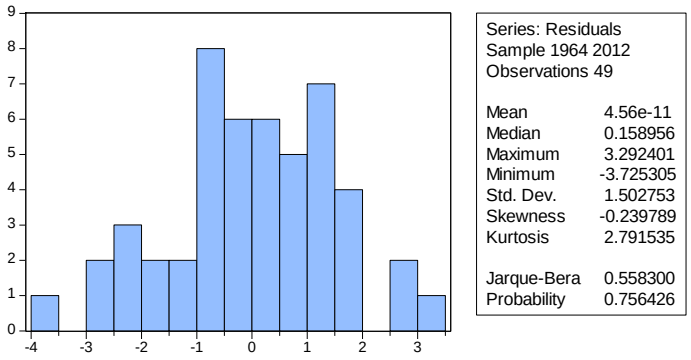
a. *Jarque-Bera test of normality* is illustrated in figure 11, and tests whether standardized residuals are normally distributed.

$H_0 : E\left(\hat{\varepsilon}_{st}\right)^3 = 0$ (*skewness*) *si* $E\left(\hat{\varepsilon}_{st}\right)^4 = 3$ (*kurtosis*)

$H_1 : E\left(\hat{\varepsilon}_{st}\right)^3 \neq 0$ *sau* $E\left(\hat{\varepsilon}_{st}\right)^4 \neq 3$

The null hypothesis of residuals normality is accepted as the p-value associated with the test has higher value than the chosen significance level.

Figure 11. *Jarque-Bera Test for Residuals*



b. *Tests for autocorrelation of residuals:* Ljung - Box Q test statistics and Breusch - Godfrey (LM).

In order to validate the model, the residuals of the estimated equation have to be a white noise process, i.e. without any correlation errors.

Ljung-Box Q test statistics: $Q_{LB} = T(T+2)\sum_{k=1}^{s} r_k^2 / (T-k)$, where T is the number of observations. The Null Hypothesys H_0: „*does not exist autocorrelation up to s*", ($r_k = 0, k = 1,...,s$, where $r_k = \frac{\gamma(k)}{\gamma(0)}$) is rejected for large values of Q or p-values less than the chosen significance level.

Figure 12. *Correlogram of Residuals*

Sample: 1964 2012
Included observations: 49

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.032	-0.032	0.0533	0.817
		2	-0.027	-0.028	0.0911	0.955
		3	0.008	0.006	0.0943	0.993
		4	-0.056	-0.057	0.2708	0.992
		5	-0.012	-0.015	0.2788	0.998
		6	0.264	0.261	4.3190	0.634
		7	0.026	0.045	4.3580	0.738
		8	-0.366	-0.384	12.521	0.129
		9	0.047	0.027	12.661	0.179
		10	-0.012	0.055	12.670	0.243
		11	0.006	0.015	12.673	0.315
		12	0.023	-0.120	12.709	0.391
		13	-0.040	-0.074	12.822	0.462
		14	-0.226	-0.025	16.453	0.286
		15	0.016	0.034	16.471	0.351
		16	0.012	-0.162	16.482	0.420
		17	0.041	0.056	16.612	0.481
		18	0.051	0.100	16.823	0.535
		19	-0.023	0.010	16.867	0.599
		20	-0.011	-0.000	16.877	0.661

The test leads to the acceptance of the null hypothesis of non autocorrelation of errors. The same conclusion is reached also if it is observed that the autocorrelation functions ACF and partial autocorrelation PACF are similar to those of a white noise process.

The Breusch-Godfrey (LM) test check also for the residuals autocorrelation. For a residuals model AR(h), the null hypothesys H_0: "*residuals are not correlated*" is rejected for that p-values asociated with F şi χ^2 statistics which are less than the chosen significance threshold. Therefore, according to the test result shown in Figure 13, the null hypothesis is accepted.

Figure 13. *Breusch-Godfrey Serial Correlation LM Test*

F-statistic	1.030304	Prob. F(2,43)	0.3655
Obs*R-squared	2.032015	Prob. Chi-Square(2)	0.3620

c. *Heteroskedasticity tests*: ARCH-LM, Breusch-Pagan-Godfrey, and White

The null hypothesis: „*homoscedasticity of residuals*" is rejected for p-values associated with F and χ^2 statistics below the chosen significance threshold. According to the test results shown in Figures 14, 15 and 16, all

statistics (F, and LM) are low and, in addition, their respective probabilities are higher by 5%, the chosen significance level. Therefore the null hypothesis is accepted, i.e. the residuals are homoscedastic.

Figure 14. *Heteroskedasticity Test: ARCH*

F-statistic	0.574626	Prob. F(1,45)	0.4524
Obs*R-squared	0.592598	Prob. Chi-Square(1)	0.4414

Figure 15. *Heteroskedasticity Test: Breusch-Pagan-Godfrey*

F-statistic	1.554971	Prob. F(1,46)	0.2187
Obs*R-squared	1.569523	Prob. Chi-Square(1)	0.2103
Scaled explained SS	1.594840	Prob. Chi-Square(1)	0.2066

Figure 16. *Heteroskedasticity Test: White*

F-statistic	2.406075	Prob. F(9,38)	0.0286
Obs*R-squared	17.42402	Prob. Chi-Square(9)	0.0425
Scaled explained SS	17.70508	Prob. Chi-Square(9)	0.0388

Step 5: Selection of the Best Model

If there are more than one ARIMA models validated for the same time series, we can choose the best of them corresponding with the minimum value of the Akaike Information Criterion (AIC) and Schwarz-Bayes Criterion (SBC) where the AIC characterize the quality estimation and SBC is a key penalty. Finally we choose the model with the lowest AIC or SBC value, that is the ARIMA(1;1;1) model:

$$HEENROLL\_LN_t - HEENROLL\_LN_{t-1} = 0.003043 + 0.520764\left(HEENROLL\_LN_{t-1} - HEENROLL\_LN_{t-2}\right) + \varepsilon_t - 0.775724\varepsilon_{t-1}$$

Breaking down the lags and first differences yields the final forecasting model that is used to forecast the HEENROLL data series.

The forecasts of the HEENROLL time series, for the next period reveal a constant tendency of the time series values for the next eight years. The mean absolute percent error, a measure often used to assess the accuracy of forecasting, and also for future comparing reasons

Conclusions

After the candidates models have been estimated and validated, and the SIC criteria above used to select the model ensures maximum efficiency, the selected model and validated is ARIMA(1,1,1):

$$HEENROLL\_LN_t - HEENROLL\_LN_{t-1} = 0.003043 + 0.520764\left(HEENROLL\_LN_{t-1} - HEENROLL\_LN_{t-2}\right) + \varepsilon_t - 0.775724\varepsilon_{t-1}$$

Here the econometric model was based on the Box-Jenkins's methodology to model the historical enrolments for fifty annually observations data, from 1963 to 2012. The time series of the percentage of the population 18 and 19 years old enrolled in higher education in USA were used to identify a forecast model, estimate its parameters, check the model's performance, and finally use it to

forecast. The empirical study revealed the best ARIMA validated model to forecast future values of the time series for the next eight years. One of the reasons for popularity of the ARIMA modeling is its success in forecasting. Unfortunately, even if the model helps us see where it's going enrolling students in the coming years, it does not help us to go behind this evolution of data. According with the senior forecaster McCalman, „forecasting methods based on univariate techniques like Box–Jenkins techniques tend to have high historical coherence but low conceptual coherence. These methods are generally very good at replicating historical movements, but provide limited insight into the causes of these movements". Because the accuracy is what matters most when it comes to forecasting, the present research will continue to compare the same data series of students' enrollment with the results of other methods that have been proposed for forecasting enrollments in the last years.

References

Boes S.and Pflaumer, P. 2005. University student enrollment forecasts analyzing structural ratios using ARIMA-methods, *Allgemeines Statistisches Archiv* 90, 253-271 Physica-Verlag 2006, ISSN 0002-6018.

Box, G. E. P., G. M. Jenkins. 1976. *Time Series Analysis: Forecasting and control*. Rev. ed. San Francisco: Holden-Day.

Chen, S. M., and Hsu C. C. 2004. New Method to Forecast Enrollments Using Fuzzy Time Series. *International Journal of Applied Science and Engineering*, 2004. 2, 3: 234-244.

Chen, C. K. 2008. An Integrated Enrollment Forecast Model. *IR Applications Volume 15, January 18, 2008* Using Advanced Tools, Techniques, and Methodologies.

Johnson, R. R. n.d. A Guide to Using EViews with *Using Econometrics: A Practical Guide*, University of San Diego. Retrieved from bit.ly/1QDHPLF.

Johnston, J. & DiNardo, J., 1997. Econometric methods, Fourth edition, McGraw-Hill.

McCalman, St. 2012. Forecasting, *National Economic Review*, Issue 67, National Institute of Economic and Industry Research (NIEIR).

Sullivan, J. and Woodall, W. H. 1994. A comparison of fuzzy forecasting and Markov modeling. *Fuzzy Sets and tems, 64: 279-293.*

Zhang J. G. 2007, Looking Ahead of the Curve: an ARIMA Modeling Approach to Enrollment Forecasting, *Association for Institutional Research*, June 4, 2007.