Title: “Forecasting student enrollment using time series models and recurrent neural networks”

Authors:

(1) Rezwanul Parvez

Research Economist

Colorado State University

Email: **rezwanul.parvez@colostate.edu**

(2) Syed Imran Ali Meerza

Assistant Professor

Arkansas Tech University

Email: **smeerza@atu.edu**

(3) Nazea H. Khan Chowdhury

Faculty of Social and Behavioral Science Program

Front Range Community College

Email: **nazea.khanchowdhury@frontrange.edu**

Selected Paper prepared for presentation for the 2021 AAEA & WAEA Joint Annual Meeting, August 1-3, 2021, Austin, TX

Date of submission:

(06/22/21)

*** This is a preliminary draft. Please do not cite or circulate without the permission of the author ***

Abstract

The key goal of this study is to project future student enrollment trends at an institutional level based on historical time-series data (1999-2020). We use available enrollment data (1999-2020) from a 4-year higher education institution at Colorado to project future trend that would lead to usage of optimal decision-making strategies needed for future planning. We propose both time-series models (ARIMA) and recurrent neural networks (RNN) models to predict future student enrollment trend. Key findings indicated the strength of recurrent neural network (RNN) -based time series forecast over a conventional time series model. This work also indicates that the performance of RNN is better in most cases relative to the auto regressive integrated moving average (ARIMA) model in predicting enrollment trends. Specifically, the forecasting performance of the RNN model is better when dealing with spring and fall semester enrollment. However, ARIMA model predicts summer enrollment better as compared to RNN models.

Keywords

Forecasting, student enrollment, time-series model, recurrent neural networks, ARIMA

1. Introduction

Academic institutions across the United States (U.S.) have experienced huge enrollment growth (24.1% increase) during the Great recession of 2007-2009 (Juszkiewicz, 2017; Mullin and Phillippe, 2009; NCES, 2010). However, enrollment trends follow a steady decline as the labor market condition gets better since post-recession. There is an ever increasing need to investigate the factors that affect student enrollment, which aid administrators attract new students and make efficient allocation of available scholarships and financial aids across academic institutions (Slim, 2018). Similarly, an accurate future enrollment trend will help make better decision regarding budget allocation, program planning, and appropriations from state legislatures. Despite the necessity, a better enrollment projection is complex due to (i) high volatility in enrollment patterns, (ii) presence of uncertainty of proper forecasting techniques, and (iii) complexity of exploring and testing factors impacting student enrollment. Here, we employ Recurrent Neural Networks (RNN) model to develop projection models based on longitudinal data on student enrollment. Additionally, we use ARIMA model as baseline to assess RNN models robustness when dealing with longitudinal data.

According to human capital theory, workers' skills and knowledge are known as function of capital gain. More years of schooling and training would bring more capital and led to economic return in form of earnings. Thus, an individual must spend more years of schooling to gain human capital in the long run (Becker, 1975). In reality, an individual tends to participate in postsecondary education and training when face labor market uncertainties like unemployment, loss of income, or layoffs. Further, according to theory of demand, enrollment decisions are a function of college's price, price of alternative educational opportunities, population changes, demographic changes, current wages and income levels, anticipated future earnings, and labor market conditions (Becker, 1990; Clotfelter, 1992; Leslie and Brinkman, 1987). Thus, it is critical to include the economic factors affecting labor market conditions into forecasting model when dealing with student population. However, a vast majority of literature have primarily focused on student population and enrollment projection tend to ignore the importance of labor market conditions (Gandy et al. 2019; Agboola and Adeyemi, 2013; Mehta, 2004; Dickey, Asher, and Tweddale, 1989; Hoenack and Pierro, 1990; Moore, Studenmund, and Slobko, 1991).

This study introduces a novel improvement of enrollment forecasting approach by employing both econometric and deep learning methods.

Conventional time series and regression analysis have been considered as the primary techniques for projection analyses in the existing literature. These models are suitable in performing the tasks of projecting values for future time period given sequence of historical records. These methods range from moving averages and Autoregressive Integrated Moving Averages (ARIMA) to Exponential Smoothing amongst others. Time series analysis is often regarded as more suitable as compared to others (subjective judgments, the ratio method, results from a cohort survival study, simulation methods, and regression analysis) if conditions are met (assumptions and longitudinal data availability) (National Center for Educational Statistics, 2013). However, time series models are limited due to its unwarranted assumptions and complexities. According to Tadayon & Iwashita (2020), ARIMA and Hidden Markov Models (HMM) are subject to restrictions for successful processing of time series data as compared to neural networks dealing with longitudinal data.

The main objectives of this work are to:

1) Identify the trend of undergraduate student enrollment

2) Propose a recurrent neural network (RNN) along with ARIMA model to determine future trends of student enrollment for higher accuracy in time-series forecasting analysis, and The analytics developed in this work are geared towards answering the following key research questions:

- What would be the future enrollment pattern of undergraduate students at an institutional level?
- Which is better method for predicting enrollment between ARIMA and RNN, considering the effect of seasonality?

Results from this research would help policymakers, educators, and researchers of higher academia make a better-informed decision. Although the direct implication may only benefit this particular institution, the approach of the model building can provide other colleges and universities with an alternative solution to undergraduate student enrollment forecasting other than intuitive estimates. Higher education institution leaders can efficiently allocate resources and do better budget management by projecting future enrollment scenario.

2. Data and Variable Selection

This study develops a comprehensive time series database to project future student enrollment trends based on past data (1999-2020) at a public institution. We include 66 data points (1 data point per term during 1999-2020) over time to discuss and project undergraduate enrollment for three future time periods (3 semesters of 2021). This sample size (66) is large enough to develop a robust ARIMA and RNN models (Chen et al, 2019).

3. Empirical Estimation Strategies

a. Test for Stationarity of Data
b. Dicky-Fuller Statistics test

Hypothesis:

- H_0: The dataset is stationary
- H_1: The dataset is not stationary

Decision: Reject the null hypothesis, H_0 if the P- value statistic is greater than 0.01 level of significance

We apply both time-series model (i.e., ARIMA) and machine learning algorithm (i.e., recurrent neural networks) to predict future enrollment. We use time series method (ARIMA) as we extracted information from historical data and forecast future behavior of the series based on past pattern. Firstly, we visualize the data and identify the data trends using programming language E-Views. Based on the trends, we build the baseline for ARIMA model as best suited.

The recurrent neural networks (RNN) method is suitable for time-series analysis. Due to the volatile pattern of enrollment data, we chose RNNs that will generate robust algorithms while modeling sequential data. To assess the prediction performance of the models, we utilize the root-mean-squared-error (RMSE).

4. Results and Discussion

Box-Jenkins Methodology for ARIMA model selection:

We select Box-Jenkins methodology to evaluate ARIMA models using the E-views 12 software. Box-Jenkins (1970) introduced a three steps method to select best model for estimating and forecasting univariate models. The three stages are:

1. **Identification** {a. stationarity b. ARIMA (determine p, d, q); here p = autoregressive order, d= integrated (difference) order, q=moving average order}

a. Stationarity (Graph, Correlogram, Formal tests [Augmented Dicky-Fuller (ADF) test, Phillips-Perron (PP) test])

If stationary: then ARMA model (p, q)

If non-stationary, then AR(I)MA model (p, d, q)

Autocorrelation function (ACF) and partial autocorrelation function (PACF)

According to Correlogram value and ADF test, enrollment data reported here are non-stationary. We fix the issue of non-stationarity of the time series by employing a number of differences (order of integration). Further, we also estimated value for Autocorrelation graphics Dickey-Fuller (Dickey and Fuller 1981), and Phillips-Perron (Phillips and Perron 1988) statistical tests to examine the order of integration of the time series (number of differences that should be applied). Thus, we select AR(I)MA model and determined the value of p, d, q as 3, 1, 3 respectively.

2. **Estimation**

We estimate the model by examining its parameters (coefficients of the autoregressive and moving average polynomials). Additionally, we preselect different architectures of ARIMA models for estimation and diagnosis as per autocorrelation and partial autocorrelation graphics of the transformed time series. For example, we tested few other potential models (e.g. ARIMA (1,0,1), ARIMA (1,1,3), ARIMA (3,0,1)) using enrollment data and chose the one that met the model selection criteria as mentioned below.

Model selection criteria:

a. Significance of the ARMA components

b. Compare Akaike, Schwartz, and Hannan-Quinn (smaller is better)
c. Maximum likelihood (bigger is better)
d. SigmaSQ: estimate of the error variance (smaller is better)

We employ ARIMA (3,1,3) model as this is a stationary and parsimonious model that fits the longitudinal data here in this case.

3. Diagnostic and Forecasting

We have our potential candidate model: ARIMA (3,1,3) for our analyses.

Requirements for a stable univariate process:

a. Residuals of the models are white noise (Ljung-Box Q statistic)
b. Check if estimated ARMA process is (covariance) stationary: AR roots should lie inside the unit circle
c. Check if the estimated ARMA process is invertible: all MA roots should lie inside the unit circle

As the above conditions are met, we can forecast with this model.

According to L-Box Q statistics, model residuals are white noise and estimated ARMA process is invertible and stationary. Thus, we chose this {ARIMA (3, 1, 3)} as our final model and forecasted enrollment for three future periods using this model. We set the model diagnosis process based on the results from the statistical tests that we run.

4a. ARIMA Baseline Forecast

According to Landeras et al., 2009 study "The ARIMA models are defined as a combination of autoregressive models and moving average models. The autoregressive models {AR(p)} generate their predictions of the values of a variable (x_t), on a number (p) of past values of the same variable (number of autoregressive delays) (x_{t-1}, x_{t-2},. . . x_{t-p}) and include a random disturbance (e_t). Similarly, the moving average models [MA(q)] develop predictions of a variable (xt) based on a number (q) of past disturbances of the same variable (prediction errors of past values) (et-$_1$, et-$_2$, . . .et-q)."

Table 1. Model Summary

Dep Var: D (undhc) **Method:** ARMA Maximum Likelihood; Sample: 1999Q2 to 2020Q2; Obs: 65				
Variable	**Coefficient**	**SE**	**t-stat**	**Prob.**
C	16.50	3079.12	0.005	0.9957
AR(3)	1.00	0.004	240.52	0.0000
MA(3)	-0.99	2.73E-05	-36588.17	0.0000
SigmaSQ	5862156	889191.9	6.592	0.0000

Statistic	**Value**	**Statistic**	**Value**
R-squared	0.718	Mean dep var	19.38
Adj.R-squared	0.704	S.D. dep var	4596.22
S.E. of regression	2499.312	Akaike info criterion	18.73
Sum Sq. Residual	3.81E+08	Schwarz criterion	18.86
Log likelihood	-604.733	Hannan-Quinn criterion	18.78
F-statistic	51.814	Durbin-Watson stat	2.57
Prob. (F-statistic)	0.000		

4b. Recurrent Neural Network (RNN) Baseline Forecast

As mentioned earlier, we also apply recurrent neural networks (RNN) to predict future enrollment. We compare our projection numbers with recently updated enrollment statistics at the academic institution. We present the actual vs predicted enrollment numbers during 1999-2020 in figure 2. For example, our prediction of this time points is 7,829 in Spring 2019, 7,647 in Spring 2020, and 2,300 in Summer 2020. The actual enrollment was 7,694 in spring 2019, 7,562 in Spring 2020, and 2,933 in Summer 2020. The data points we estimated stayed within the prediction interval.

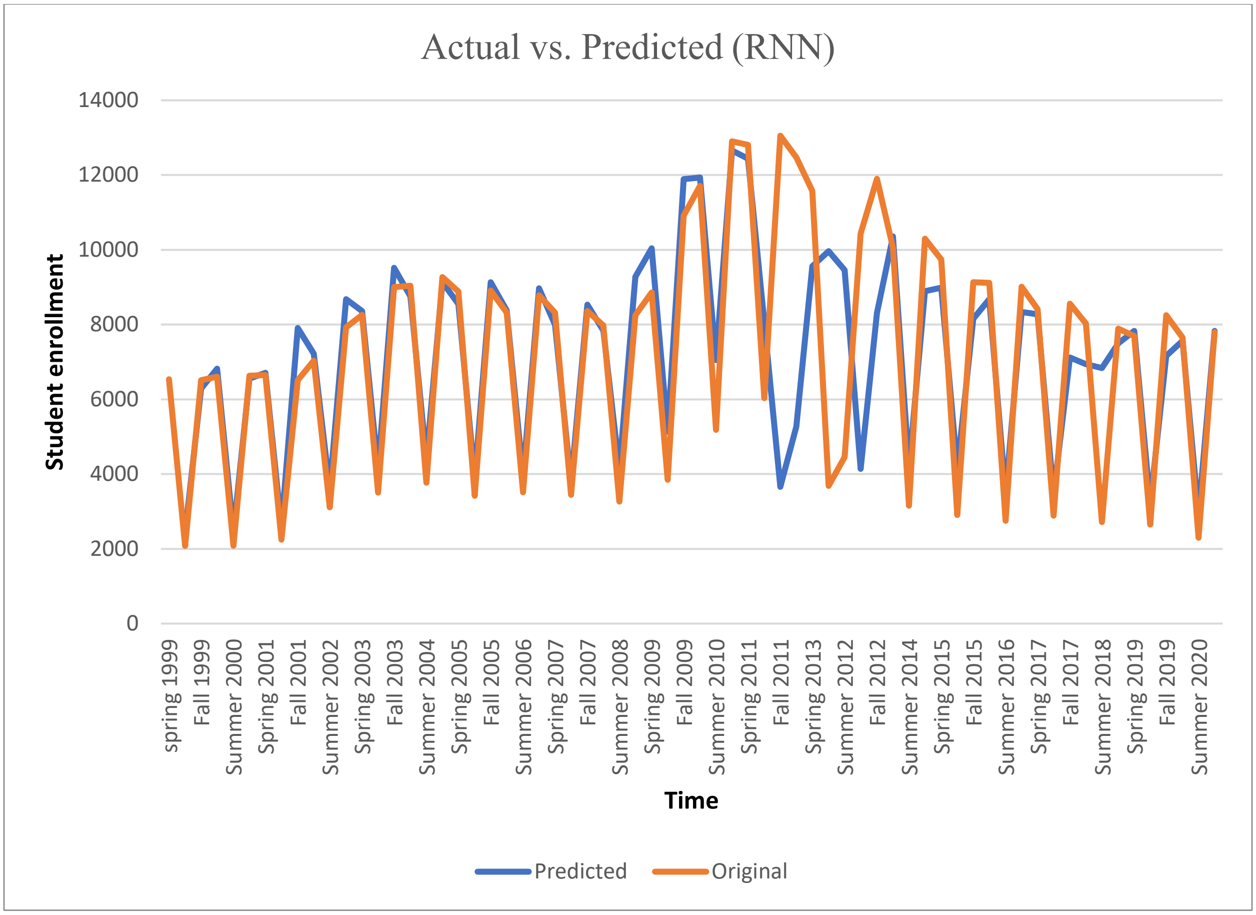

Fig 1. Actual vs Predicted Value (RNN)

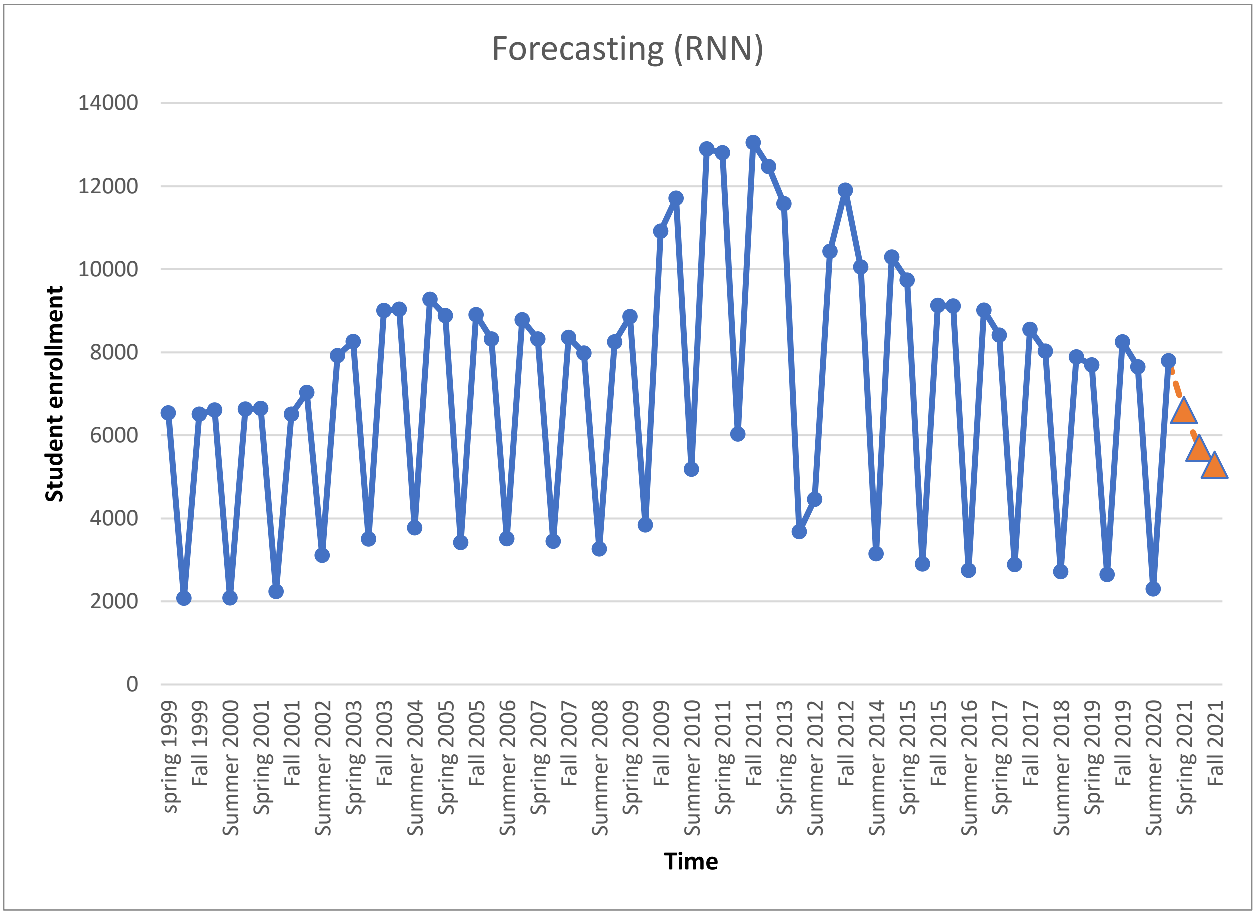

Fig 2. Future Forecasting Trends

According to Fig 2, our estimated future values are 6,609 for Spring of 2021, 5,696 for Summer of 2021 and 5286 for Fall of 2021, respectively (as per RNN model).

Table. 2 Comparison of ARIMA & RNN model performance

Year	**Quarter**	**Forecasted value (ARIMA)**	**Forecasted value (RNN)**
2021	Q1-Spring	7869.00	6609.50
2021	Q2-summer	3434.98	5696.77
2021	Q3-Fall	7897.99	5286.39

Three ARIMA models and one ANNs were employed as these models showed promise during model identification, estimation, and diagnostic processing stages.

RNNs models performed better to estimate future number for spring and fall semester enrollment. However, ARIMA model perform better for predicting summer enrollment number as compared to RNN models. As per ARIMA model, our estimated future values are 7,869 for Spring of 2021, 3,434 for Summer of 2021 and 7,897 for Fall of 2021, respectively.

5. Conclusion

This study is an attempt to examine 22-year historical records of undergraduate student enrollment data and identify future enrollment pattern using both time series models and RNN models. Student enrollment data is volatile in nature and shows irregular pattern over time.

Higher educational institutions should adopt robust memory based RNN model as compared to conventional time series models to achieve better accuracy for enrollment forecasting. RNN projection techniques seems to have lower errors in the forecast as compared to traditional timeseries models. We observe higher differences in case of quarterly time series with more variability between Fall and Summer enrollment numbers. Due to this volatile nature of enrollment data and presence of seasonality, we recommend the need for further evaluation studies of other types of ARIMA models, artificial neural networks, and hybrid ARIMA-artificial neural network models to forecast student enrollment. Inclusion of local labor market analytics data may enable to capture enrollment variation pattern over time to achieve better enrollment accuracy.

References

1. Agboola, B. M., & Adeyemi, J. (2013). Projecting Enrollment for Effective Academic Staff Planning in Nigerian Universities. Educational Planning, 21(1).
2. Becker, G. S. (1975). Human capital: A theoretical and empirical analysis, with special reference to education. National Bureau of Economic Research, New York, NY.
3. Becker, W. (1990). The demand for higher education. In S. Hoenack & E. Collings (Eds.), The economics of American Universities: management, operations, and fiscal environment (pp. 155-188). New York: SUNY.
4. Box, G. E. P., and G. M. Jenkins. (1970). Time Series Analysis Forecasting and Control. San Francisco: Holden-Day.
5. Chen, Yu April, Ran Li and Linda Serra Hagedorn. (2019). Undergraduate International Student Enrollment Forecasting Model: An Application of Time Series Analysis. Journal of International Students, Volume 9, Issue 1, 242–261.
6. Clotfelter, C. T. (1992). Explaining the demand. In C. T. Clotfelter, R. G. Ehrenberg, M. Getz, & J. J. Siegfried (Eds.), Economic challenges in higher education (pp. 59-88). Chicago: The University of Chicago Press.
7. Dickey, A. K., Asher, E. J. Jr., & Tweddale, R. B. (1989). Projecting headcount and credit hour enrollment by age group, gender, and degree level. Research in Higher Education, 30(1), 1–19.
8. Dickey DA, Fuller WA. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49: 1057–1072.
9. Gandy, R., Lynne, C., Andrew, L., Daniel, K., & Sherry, K., (2019). Enrollment Projection using Markov Chains: Detecting leaky pipes and the bulge in the Boa. The AIR Professional File.
10. Hoenack, S. A., & Pierro, D. J. (1990). An econometric model of a public university's income and enrollment. Journal of Economic Behavior and Organization, 14(3), 403–423.
11. Juszkiewicz, J. (2017, November). Trends in Community College Enrollment and Completion Data, 2017. Washington, DC: American Association of Community Colleges.

12. G. Landeras, A. Ortiz-Barredo, J.J. Lopez (2009). “Forecasting weekly evapotranspiration with ARIMA and artificial neural network models” ASCE Journal Of Irrigation and Drainage Engineering., 135 (3), pp. 323-334.
13. Larry L. Leslie and Paul T. Brinkman. (1987). Student Price Response in Higher Education: The Student Demand Studies. The Journal of Higher Education Vol. 58, No. 2 (Mar. - Apr., 1987), pp. 181-204.
14. Mullin, C., & Phillippe, K. (2009). Community college enrollment surge. AACC policy brief series. Washington, D.C.: American Association of community colleges.
15. Mehta, A. C. (2004.) Projections of population, enrollment and teachers.
16. Moore, R. L., Studenmund, A. H., & Slobko, T. (1991). The effect of the financial aid package on the choice of a selective college. Economics of Education Review, 10(4), 311–321.
17. P.C.B. Phillips, P. Perron (1988). Testing for a unit root in time series regression. Biometrika, 75, pp. 335-346.
18. Slim, A., Hush, D., Ojah, T., & Babbitt, T. (2018). Predicting Student Enrollment Based on Student and College Characteristics. International Educational Data Mining Society.
19. Tadayon, M., & Iwashita, Y. (2020). Comprehensive analysis of time series forecasting using neural networks. arXiv preprint arXiv:2001.09547.
20. National Center for Education Statistics (NCES), (2010; 2013; 2019). Trends in Undergraduate Nonfederal Grant and Scholarship Aid by Demographic and Enrollment Characteristics: Selected Years, 2003–04 to 2015–16. Available at https://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=2019486